Prediction of poverty in Latin America: model selection for the 2000s census round

Pier Francesco De Maria*
UNICAMP, Campinas, Brazil – dpierf@gmail.com

Abstract

Poverty in Latin America is an important question to be solved nowadays; a common pattern can be defined for the whole continent, in order to design correctly poverty indices and public policies. Data were collected from IPUMS International for 9 countries of Latin America; the method combines logistic regressions with ROC curves, which utility is defining a best choice model in the continent. The variables of the best models could be considered as appropriated to elaborate multidimensional poverty indices. Key results show that exists a common poverty pattern in the continent; the confounding factors assume a unique tendency for all the countries. Lastly, the paper concludes that variables for the "educational-working" and the "facilities and utilities" models can be used with the aim of elaborating a MPI, because of them representativeness.

Keywords: binary logistic regression; ROC curve; model comparison.

1. Introduction

Poverty in Latin America is an important theme to be analysed and studied, for public policies and definition of common benchmarks. Moreover, know more about what poverty is and how to measure this phenomenon is a relevant question, especially in the actual context of reduction of poverty and inequality in Latin America. Discussing poverty in a single country is quite simple, because one can take the data of one census or Survey and analyse the dimension and the depth of poverty. A bit more complicated is to study poor people among decades, using more than one census or yearly surveys, like PNAD (National Household Sample Survey) in Brazil, because the questionnaire changes year by year, and this can make comparisons difficult.

However, when one is interested in discuss and compare poverty in a whole continent, more problems appear. Some of them can be summarized as follows: 1) the spatial-analysis-unit problem; 2) the dimension of each sample; 3) the techniques of sample construction; 4) the year of each survey or census (a "time difference" problem); and 5) the questionnaire problem. Those problems, indeed, can make the analysis of poverty and the comparisons between countries difficult (in order to define a common benchmark). Though, one cannot simply reject the use of more than one country census because of these limits. To overcome at least two of the questions above (the fourth and the fifth), a solution is create "poverty models", using groups of variables, for each country. Then, one can compare them based on their prediction power, using ROC (Receiver Operating Characteristic) curve. The last step is comparing selected countries, in order to verify which model is the most relevant.

The aim of this paper is measure, compare and describe poverty in Latin America, using data for the "2000s census round" (taken between 2001 and 2011) of 9 selected countries. The data were collected from IPUMS (Integrated Public Use Microdata Series) International, selecting a sample of 5% for those countries. The central hypothesis of the paper is a common preferable model for the majority of the selected countries; moreover, another important hypothesis is the differential role that confounding factors (gender, race, marital status and household location) play in each country. The main model is a binary logistic regression for each country, without corrections for heteroskedasticity or autocorrelation. Each country is modelled using four poverty-types: 1) the socio-demographic characteristics model; 2) the educational-working model; 3) the household features model; and 4) the household's facilities and utilities model.

This paper is divided into four sections, including this introduction. The next section presents the data for Latin America, exhibits the methods of analysis and describes each problem faced in this research. The third one shows the main results and the conclusions that can be retrieved from ROC analysis for each country. Lastly, the fourth section analyses the results obtained and concludes about the contributions of the paper and the power of the method to address the purposed problem.

2. Data and methods

In order to analyse poverty in Latin America, the only available data with representativeness among years and regions, which is conducted in all the countries, is the census. Each country carries out its census approximately every 10 years; census is usually performed with one or two questionnaires: when there are two, one is shorter (for the universe) and the other is wider (for a sample between 5 and 10% of the population). When a research is made within a country, it is simpler to make comparisons of populations, because everyone has answered the same questionnaire: in this case, spatial and temporal comparisons are easier to be made. Otherwise, when one compares more than one country, the same questionnaire is not applicable to all the populations; indeed, the same variables used in a country could not be available in another.

To solve this problem, a possible solution is create more than one model, notwithstanding a few variables could not be available in some censuses. The main idea behind the division is that the variables can be grouped by type (common characteristics); even in the case that one or two variables were absent, estimation for the model can be made, because the core of the model is still there. Furthermore, separate variables into groups allows analyse and compare different dimensions of poverty. The analysis will be made with binary logistic regression and the comparatives are going to be done using ROC (Receiver Operating Characteristic) curve. The goals of this two-step study are the following: 1) verify which model best describes poverty in each country; 2) describe the impact of the variables in the probability of being poor; and 3) test the hypothesis of a common poverty-pattern in whole Latin America.

The logistic regression (more often known as "logit model") is "often more convenient than the probit model and has gained popularity in econometrics" (Chow, 1983: 255). This model is very useful when the dependent variable is discrete and, specifically in the case of binary logistic, can represent the probability of a person being or not in the control group. In this paper, the "control group" represents the poor people. Algebraically, one can represent the logit model like is done by Greene (2003: 667) as follows, making the well-known logarithmic transformation on our behalf.

$$Prob(Y = 1|\mathbf{x}) = \frac{e^{\mathbf{x}'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}'\boldsymbol{\beta}}} = \Lambda\left(e^{\mathbf{x}'\boldsymbol{\beta}}\right) \longrightarrow \ln\left(\frac{Prob(Y = 1|\mathbf{x})}{1 - Prob(Y = 1|\mathbf{x})}\right) = \mathbf{x}'\boldsymbol{\beta} \qquad \textbf{(1)}$$

Since there is no evidence of autocorrelation or heteroskedasticity in our data, the traditional method of maximum likelihood estimators (MLEs) is developed, as presented in Chow (1983: 256-7).

As it is mentioned in the introduction, four models are created[1]. The first one (about socio-demographic characteristics), includes: number of people in the household (*persons*); household type (*hhtype*); number of families living together (*nfams*); and birthplace of the household head (*nativty*). The second model (about education and working status of the head of household) uses the following elements: educational attainment (*edattan*); employment status (*empstat*); class of worker (*classwk*); and age in years (*age*). The third one (that analyses some household characteristics) uses variables about: electricity (*electrc*); water supply (*watsup*); sewage (*sewage*); and cooking fuel (*fuelck*). Lastly, the fourth model (about consumer goods) evaluates: telephone availability (*phone*); presence of a computer (*computr*); television set (*tv*); refrigerator (*refrig*); car ownership (*autos*); and the toilet-type (*toilet*). If one or more of these variable was not available, they would be not considered.

For the dependent variable Y, the idea of poverty as a relative phenomenon should be used, because the approach of "having less than others in society" (Hagenaars & De Vos, 1988: 214) in an unequal continent like Latin America is extremely appropriate. This method is the most practical and the simplest, in order to define who are the poor in more than one country. However, census data for Latin America do not provide the total income received monthly for all the countries; so, we cannot simply calculate a relative measure of poverty, like a fraction of median income in each country.

In order to overtake this problem, a proxy should be elected, considering two points: the presence of the variable in all the censuses; and the similarity with income or wealth. A variable that fulfil both requirements is the ownership of dwelling (*ownrshp*), under the assumption that poorer people do not own the housing or, at least, do not live in a rented house (for example, it is provided by

[1] The variables for each model are described and the IPUMS codes are shown into brackets in italic.

employer or it is an irregular dwelling, like Brazilian *favelas*). With the aim of create a dummy for Y, we categorize the response-codes of *ownrshp* (settled by IPUMS) as below, identifying the poor with 1 and the non-poor with 0. In equation (1), P_i refers to the probability of a being poor.

$$Y_i = \begin{cases} 0, & \text{if } ownrshp = \{100, \dots, 239\} \\ 1, & \text{if } ownrshp = \{240, \dots, 999\} \end{cases} \implies P_i = P(Y_i = 1) \qquad \textbf{(2)}$$

The main codes are: 110 (owned, already paid); 120 (owned, still paying); 210 (renting); 240 (occupied *de facto*); 250 (free or usufruct); and 290 (other not-owned cases). Table 1 shows this fact for the 9 selected countries.

Table 1 – Number of households by ownership of dwelling and country

Y_i	Type	BOL	BRA	CHI	COL	ECU	MEX	PER	URY	VEN
0	**Owned**	132.1	2136.8	301.2	584.2	203.3	2402.3	481.3	66.1	416.9
	Renting	32.5	454.4	72.9	272.4	81.8	253.8	102.3	23.6	64.6
1	**Provided**	7.7	91.9	19.0	.	6.0	.	.	2.1	.
	Free	.	170.6	21.5	112.0	49.0	.	24.4	.	37.9
	Other	2.2	16.8	.	24.3	.	231.1	31.5	.	15.7
%	**$Y_i = 0$**	**94.3**	**90.3**	**90.2**	**86.3**	**83.8**	**92.0**	**91.3**	**97.7**	**90.0**
	$Y_i = 1$	**5.7**	**9.7**	**9.8**	**13.7**	**16.2**	**8.0**	**8.7**	**2.3**	**10.0**

Source: Integrated Public Use Microdata Series, International: Version 6.3 (2014).
Note: data for household in thousands. An empty cell indicates no available data.

Data come from the censuses of 9 countries in Latin America: Bolivia; Brazil; Chile; Colombia; Ecuador; Mexico; Peru; Uruguay; and Venezuela. The censuses were conducted between 2001 and 2011, and the data were extracted from IPUMS International, considering a sample of 5% for all the countries[2]. The variables selected above are previously harmonized and comparable among countries, with the restriction of availability of the questions into the census' questionnaire. The four models described above include (when available) four confounding factors, which are strictly necessary in order to interpret correctly the coefficients' estimates: gender (*sex*); race or colour (*race*); marital status (*marst*); and household location (*urban*). Those variables were chosen because of the well-known differences, described in the literature: between males and females household head (e.g., Suter & Miller, 1973); about the urban-rural inequality; upon living conditions related to marital status (e.g., Waite, 1995); and considering race inequality (e.g., Wright, 1978).

After executing the logistic regressions, the next step is analysing the models with ROC curves. The ROC (Receiver Operating Characteristics) curve is used when we aim to "measure the quality of diagnostic information and of diagnostic decisions in a meaningful way" (Metz, 1978: 283). In this paper, the main idea of using ROC curves is comparing models in aspects like sensitivity and specificity. To do that, one compares what the model states about an observation (if it is positive or negative) with the reality. Then, one can have: true positives (TP); true negatives (TN); false positives (FP); and false negatives (FN)[3].

In this paper, the positive observation is $Y = 1$; in other words, when one is identified as poor, that is the positive observation. In order to compare four models in 9 countries it is also important to use a unique measure of performance. A common method is the area under the ROC curve (AUC), which varies between 0 and 1. Hanley & McNeil (1982: 30) states that this area "corresponds to the probability of correctly identifying" an observation, in our case, as poor or not poor; generally, as Fawcett (2006: 868) states, "because random guessing produces the diagonal line between (0,0) and (1,1), which has an area of 0.5, no realistic classifier should have an AUC less than 0.5".

So, the aim of this paper is define which model best predicts poverty in Latin America. The use of logistic regression is justified in order to obtain a binary predictor that can be analysed (in terms

[2] 8 out of 9 countries have a sample of 10%. These samples were reduced in order to have the same size of the smallest sample available (Brazil, 5%) for all the countries. Moreover, this reduction does not affect the statistic estimation.
[3] The principal measures of model performance are summarized by Fawcett (2006: 862).

of precision, accuracy, etc.) and evaluated with a ROC curve. The use of three different models has the objective of comparing the prediction power of characteristics of poverty. Finally, the comparison of countries in Latin America is useful, in order to verify if one model is the best choice for the continent.

3. Results and discussion

The four models of logistic regressions have, when data for each country are available, four confounding factors: sex, gender, marital status and household location. Analysing the outputted odds ratios for each country, the general results about these variables are the following:

- Widowed people are less likely to be poor, compared to single and married ones;
- Comparing married and single people, the married ones are less likely to be poor;
- Living in rural areas increases a lot the chances of not owning the dwelling;
- Men are more likely to be poor, if compared to women that are head of household;
- Analysing races, it seems that white people are less likely to be poor than black ones.

These results confirm that the use of confounding factors are strictly necessary, in order to evaluate correctly the odds for the other characteristics. But the differential role of these variables in each country could not be confirmed, because the odds ratios for these ones are similar (indeed, a confounding factor has the same effect in the whole continent). The next step is analyse, without considering the confounding factors (already described), the four models elaborated.

The "socio-demographic characteristics" model shows us that a bigger family is less likely to be poor (possibly because more people are related to a greater income); otherwise, a greater number of family living together is, probably, the clearest sign of poverty in this model. Analysing the household type, it seems to be clear that the one-person type is likely to be poorer; moreover, couples without children are generally less poor than couples with one or more kids. Lastly, considering the nativity status, native-born people are much more likely to be poor than foreign-born ones.

The "educational-working status" model points out some interesting characteristics. About the educational attainment, it is very evident that a more educated household's head has a lower chance of being poor; but the most striking difference is between a head with an university degree and one with only a secondary school completed. Regarding the working class, self-employed are much less likely to be poor, comparing them with salary workers and unpaid workers; additionally, salary workers have a greater probability of being poor than the unpaid ones. At least, a curious result appears when one compares employment statuses: the inactive people are less likely to be poor than employed and much less likely than unemployed one. This result may be an effect of a composition of some factors, as: receive a pension; and have built or paid the dwelling in the past.

The third and fourth models (the "household characteristics" one and the "facilities and utilities" one, respectively) indicate that the presence of facilities/utilities and a better household structure are strictly associated to the ownership of the dwelling. These and more results are show in the Table A.1 in appendix, were coefficients of odds ratio are compared for the independent variables by country. For the confounding factors and the constant, an arithmetic mean is calculated, in order to exhibit the average odds ratio for each situation.

The predicted values for each model and country were used in order to draw ROC curves. These curves were drawn for the 9 countries and the whole Latin America. The most important results can be resumed using Table 2; an easy manner to resume ROC analysis is the AUC index presented in the previous section. The most efficient models are the second (about educational and working statuses) and the fourth (about household facilities and utilities), with a slightly better performance of the latter over the former. This better performance can probably be associated to the dichotomous nature of the variables of that model.

Comparing the results among the countries, with a simple average of the 4 models AUC, the best performance is for Colombia, which has also the fewest missing variables for this research. At means, the models analysed here have an average performance of 0.67 (for Latin America, with a worse result for Peru, below 0.60); this result, for real data in social science, can be considered as a satisfactory performance.

Table 2 – Area under ROC curve for each model and country and for Latin America

	BOL	BRA	CHI	COL	ECU	MEX	PER	URY	VEN	LAC
Model 1	0.600	0.642	0.654	0.698	0.601	0.575	0.556	0.615	0.617	**0.646**
Model 2	0.662	0.672	0.680	0.743	0.651	0.630	0.591	0.675	0.708	**0.679**
Model 3	0.588	0.642	0.660	0.764	0.601	0.579	0.630	0.608	0.636	**0.658**
Model 4	0.612	0.672	0.692	0.779	0.619	0.633	0.599	0.700	0.690	**0.683**
Average	**0.616**	**0.657**	**0.672**	**0.746**	**0.618**	**0.604**	**0.594**	**0.650**	**0.663**	**0.667**

Source: results obtained using data from IPUMS, International: Version 6.3 (2014).

The hypothesis of a "common preferable model" is not rejected, because 5 out of 9 countries best predict our poverty proxy with the fourth model. In other words, knowing which are the facilities and the utilities of a dwelling makes simpler to describe the poverty profile of Latin America. This means that a simple poverty index can be defined using only data from Model 4; if one also wants to know about the population characteristics, in order to elaborate a more complex poverty index, this task can be accomplished using data from educational attainment and/or working status (Model 2).

4. Conclusions

This paper had the aim of discussing manners of analysing poverty in Latin America. The objectives were: measure poverty in the continent using a proxy, like the ownership of dwelling; compare models based on specific characteristic, present in census data made available from IPUMS; and define a poverty common descriptor for the whole continent. The data shows us that a common pattern of poverty can be outlined, even if, in some cases, not all variables where available. Moreover, the confounding factors of the paper have the same impact among all the selected countries, which also demonstrates that poverty in the continent is homogenous for these basic characteristics. In other words, gender inequalities, marital and urban/rural status and race/colour differences are similar in all the countries of Latin America, outlining a structural contrast between socioeconomic categories. Ultimately, even if some data limitation have slightly reduced the comparisons between countries (and the possibility of measure poverty), the explanatory power of the models was satisfactory. This paper concludes, thereby, that poverty in Latin America is a complex but similar (across countries) phenomenon, with the greatest difference concerning the stage of this problem in each country.

Acknowledgments

The author wishes to acknowledge the statistical offices that provided the underlying data making this research possible: National Institute of Statistics, Bolivia; Institute of Geography and Statistics, Brazil; National Institute of Statistics, Chile; National Administrative Department of Statistics, Colombia; National Institute of Statistics and Censuses, Ecuador; National Institute of Statistics, Geography, and Informatics, Mexico; National Institute of Statistics and Informatics, Peru; National Institute of Statistics, Uruguay; and National Institute of Statistics, Venezuela.

References

Chow, G. (1983). *Econometrics*. New York (NY), USA. McGraw-Hill Book Company.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, **27**, 861-874.

Greene, W.H. (2003). *Econometric Analysis*. Upper Saddle River (NJ), USA. Prentice Hall.

Hagenaars, A. & De Vos, K. (1988). The Definition and Measurement of Poverty. *The Journal of Human Resources*, **23**(2), 211-221.

Hanley, J.A. & McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**(1), 29-36.

Metz, C.E. (1978). Basic Principles of ROC analysis. *Seminars in Nuclear Medicine*, **8**(4), 283-298.

Minnesota Population Center (2014). *Integrated Public Use Microdata Series, International:* Version 6.3 [Machine-readable database]. Minneapolis: University of Minnesota.

Suter, L.E. & Miller, H.P. (1973). Income Differences Between Men and Career Women. *American Journal of Sociology*, **78**(4), 962-974.

Waite, L.J. (1995). Does Marriage Matter? *Demography*, **32**(4), 483-507.

Wright, E.O. (1978). Race, Class and Income Inequality. *Am. Journal of Sociology*, **83**(6), 1368-1397.

APPENDIX

Table A.1 – Odds ratio for the logistic regression by model, variable and country

			BOL	BRA	CHI	COL	ECU	MEX	PER	URY	VEN
Constant (average of the 4 models)			0.07	0.05	0.06	0.07	0.16	0.08	0.14	0.52	0.08
Confounding factors (average of the 4 models)	**Race or colour**	White	.	0.93	.	1.55	0.81	.	.	1.02	.
		Black	.	1.06	.	1.45	1.03	.	.	1.31	.
		Indigenous	.	0.66	.	0.76	0.58	.	.	1.20	.
		Others	Reference								
	Marital status	Single/never married	1.59	1.36	1.72	1.35	1.37	1.60	1.31	1.11	2.70
		Married/in union	1.54	1.25	1.61	1.38	1.41	1.15	1.34	1.22	2.12
		Separated/divorced	1.60	1.50	1.82	1.30	1.43	1.69	1.54	1.30	2.06
		Widowed	Reference								
	Sex	Male	1.14	1.19	1.22	0.96	1.06	1.03	1.03	1.08	1.45
	Status	Rural	0.67	3.05	3.34	2.79	1.65	.	0.91	.	1.71
Model 1 Socio-demographic	**Number of persons**		0.97	1.00	1.02	0.93	0.98	0.95	0.98	1.15	1.01
	Number of families		1.23	1.16	1.00*	1.11	1.10	1.29	1.08	0.92	1.25
	Household classification	One-person	Reference								
		Couple, no children	0.77	0.72	0.58	1.19	0.85	0.71	0.70	0.51	0.70
		Couple, with children	1.13	0.83	0.78	1.38	1.20	0.85	0.94	0.62	0.64
		Single-parent family	0.95	0.77	0.75	0.87	1.02	0.82	0.91	0.74	0.73
		Extended family	0.76	0.55	0.56	0.83	0.81	0.66	0.73	0.40	0.44
		Composite	0.62	0.53	0.51	0.98	0.96	0.43	0.55	0.51	0.54
		Non-family	1.07	0.73	0.79	0.83	0.82	0.35	0.86	0.62	0.90
	Nativity status	Native-born	1.44	1.28	1.38	1.62	1.18	1.29	1.92	1.43	0.86
Model 2 Educational-working status	**Age**		0.98	0.98	0.98	0.98	0.97	0.99	0.99	0.98	0.96
	Educational attainment	Less than primary	1.35	2.32	2.42	2.25	1.92	2.11	1.97	6.20	3.54
		Primary completed	1.56	2.09	2.09	1.86	1.57	2.09	1.87	4.13	2.61
		Secondary completed	1.60	1.55	1.66	1.39	1.34	1.51	1.58	1.90	1.64
		University completed	Reference								
	Class of worker	Self-employed	0.98*	0.77	0.65	0.74	0.81	1.30	0.98	.	0.71
		Wage/salary worker	2.27	1.40	1.06	1.07	1.13	1.80	1.29	.	1.11
		Unpaid worker	Reference								
	Employment status	Employed	1.02	1.02	1.16	14.74	0.86	1.42	1.01	0.95	0.80
		Unemployed	1.45	0.91	1.12	4.74	0.97	1.18	1.14	1.24	0.92
		Inactive	Reference								
Model 3 Household features	**Cooking fuel**	None	2.55	.	.	12.81	2.00	6.72	1.69	2.01	.
		Electricity	1.30	.	.	0.58	0.90	1.78	1.30	0.80	1.01*
		Petroleum or gas	1.59	.	.	0.56	1.26	1.24	1.45	0.42	1.34
		Other	Reference								
	Water supply	Yes	0.93	1.16	1.31	0.75	0.82	0.80	0.61	0.50	0.63
	Sewage	Connected	0.85	0.78	0.50	.	0.78	0.85	0.54	0.47	0.61
	Electricity	Yes	1.01*	0.93	0.56	1.10	0.72	0.68	0.57	0.43	0.93
Model 4 Facilities/utilities	**Television set**	No	0.99*	1.10	1.09	1.38	.	0.86	1.06	1.43	1.28
	Automobile	No	1.61	1.28	1.11	1.57	.	1.45	.	1.52	1.86
	Computer	No	.	1.21	1.24	1.42	1.34	1.33	1.48	1.44	1.56
	Refrigerator	No	1.30	1.07	1.23	1.25	.	1.12	1.19	1.57	1.69
	Telephone	No	1.29	1.51	1.76	1.37	1.56	1.36	1.41	2.00	1.45
	Toilet	No	0.74	.	2.67	0.95	1.24	1.29	1.57	2.40	1.05

Source: results obtained using data from IPUMS, International: Version 6.3 (2014).
Notes: significance at 5% level. Asterisks indicate insignificant coefficients. Points indicate inexistent variables.