Heteroskedastic Generalized Linear Models to explore income inequality in Latin American countries

Gilbert Brenes-Camacho*
School of Statistics, Universidad de Costa Rica, San José, Costa Rica– gbrenes@ccp.ucr.ac.cr

Abstract

The heteroskedastic generalized linear model (HGLM) is a generalized linear model (GLM) where the conditional variance of a dependent variable yi is a function of a vector of covariates xi. This paper aims to analyze differential income inequality among groups in Brazil, Colombia, and Uruguay –all countries in Latin America, a region with high income inequality–, by modeling log-income variance. Heteroskedastic models can be useful to understand the characteristics of economic inequality in the region, and highlight the variables that can be subject to public policies in order to improve income distribution. We use household surveys datasets produced by these countries' National Statistical Offices, and standardized by the Luxembourg Income Study (LIS) project to enhance comparability. The analysis starts with a Variance-Equation-only model for each country to determine between-groups differences in income variability. A Mean-and-Variance-Equations model is also estimated to control for between-groups differences in mean income. We compute a ratio of coefficients of both models to determine if differences in income inequality would diminish after controlling for mean differences. In all three countries, income inequality is greater among older people, women, the self-employed labor force, high-skilled workers, and people working in the agricultural sector. After controlling for mean-income differentials, the differences in income variability increases between men and women, and between dependent and independent workers.

Keywords: Heteroskedastic linear models; income inequality; log-income variance.

1. Introduction

The heteroskedastic generalized linear model (HGLM) is a generalized linear model (GLM) where the conditional variance of a dependent variable yi is a function of a vector of covariates xi. Western and Bloome (2009) have proposed to analyze socio-economic inequality using HGLM. The Latin American region includes some of the countries with the highest income inequality in the world. This paper aims to estimate a model of income inequality for several Latin American countries; in the model, the conditional mean and the conditional residual variance are regressed on a set of covariates in order to estimate equations that can be used to construct scenarios of changes in income inequality.

2. Heteroskedastic Generalized Linear Models

Let a GLM be defined as

$$g\big(E(y_i|\boldsymbol{x_i})\big) = \boldsymbol{x_i}'\boldsymbol{\theta}$$

where y_i is the dependent variable, x_i is the vector of covariates, g(.) is the link function, and θ is the vector of regression coefficients (Nelder and Wedderburn, 1972). The conditional distribution of y_i given x_i implies a specific relationship between the conditional mean and the conditional variance.

Following Smyth (1989), the relationship between the conditional mean and the conditional variance can be expressed as:

$$Var(y_i|x_i) = \phi_i w_i^{-1} \nu(E(y_i|\boldsymbol{x_i}))$$

where ν(.) is a non-negative scalar function defined by the conditional probability of y_i given x_i; w_i is a weight factor, and ϕ_i is a dispersion parameter relative to observation i.

In a HGLM, the dispersion parameter can be modeled as a function of a set of covariates z_i:

$$h(\phi_i) = \boldsymbol{z_i}'\boldsymbol{\gamma}$$

where γ is a vector of coefficients, and h(.) is the link function that defines the relationship between the covariates and the dispersion parameter (Smyth, 1989). Smyth (1989) explains that γ can be estimated considering a model where the deviance components of the conditional mean model are functions of the covariates z_i, given that the mean and the dispersion parameter are orthogonal to each other.

There are four proposed methods or approaches for estimating HGLM: a two-stage method, Maximum Likelihood Estimation (MLE), Restricted Maximum Likelihood Estimation (RMLE), and Bayesian Estimation. Zheng et al. (2013) compare the advantages and disadvantages of each method. Given that we use these models in large sample datasets, we are going to use MLE. It is worth noticing that HGLM are also known as Double Hierarchical Generalized Linear Models and Generalized Additive Models for location, scale, and shape (Zheng et al., 2013).

3. Analysis of income inequality using HGLM

The Gini index is probably the most popular measure of income inequality. Its extensive use is linked to its desirable properties: scale invariance and sensitivity to the principle of transfers. An inequality measure is scale invariant if its value remain the same after all the income values are multiplied by a constant; a measure is sensitive to the principle of transfers if it represents a more equal distribution when certain amount of income from a richer person is transferred to a poorer person. Theil's index and the Coefficient of Variation share these properties (Allison, 1971). The variance of the natural logarithm of income (L) is also scale invariant, but it does not satisfy the principle of transfers properly at large incomes. Nonetheless, the violation to this principle does not greatly affect its interpretability in most empirical contexts (Western and Bloome, 2009). Moreover, under the assumption that the income distribution is lognormal –that is, log-income has a Gaussian distribution–, L can be transformed into the Gini and Theil's indices and into the Coefficient of Variation (Allison, 1978). Allison (1978) also argues that, if income inequality measures are computed from a sample, it is better to compute confidence intervals to L, and then calculate the confidence limits to the other measures, rather than trying to compute them directly, given the inferential properties of L.

Considering the advantages and disadvantages of L, a HGLM where log-income is the dependent variable can be used to analyze income inequality as a function of several covariates (Western and Bloome, 2009; Zheng et al., 2013). If y_i refers to income, then HGLM results can be expressed as a system of equations:

$$E(\log(y_i)|\boldsymbol{x_i}) = \boldsymbol{x_i}'\boldsymbol{\theta} \qquad \textbf{(Mean equation)}$$

$$\widehat{Var}(\log(y_i)|\boldsymbol{x_i}) = exp(\boldsymbol{x_i}'\boldsymbol{\gamma}) \qquad \textbf{(Variance equation)}$$

The natural logarithm is used as a link function for the variance equation in order to assure that the variance is always positive.

If no covariates are added to the mean equation (this the Variance-Equation-only model), the $\boldsymbol{\gamma}$ coefficients represent the differences in income inequality between the groups defined by the levels of $\mathbf{x_i}$. The results would be equivalent to computing inequality measures (Gini Index, Theil´s measure, etc.) for different subgroups. Aside from parsimony, the HGLM has the advantage that it is possible to contrast the hypothesis of equal variances between groups.

If no covariates are added to the variance equation (a Mean-Equation-only model), the predicted variance would be equal to the mean error variance of a typical GLM; in a Gaussian model, the

predicted variance would be equal to the Mean Square Error (MSE). Therefore, if covariates are added to both equations, the variance equation models the residual variance after controlling for the conditional mean of log-income given the set of covariates (Western and Bloome, 2009).

Let γ_j and $\gamma_{j|\theta}$ be the regression coefficients of covariate x_j in the variance equation when no covariates are added to the mean equation and in the variance equation when covariates are added, respectively, then:

$$\Delta_{\gamma_j|\theta} = {}^{\gamma_{j|\theta}}/_{\gamma_j}$$

the ratio $\Delta_{\gamma_j|\theta}$ expresses the relative change in residual income inequality differentials when differences in mean income across subgroups (covariates) are taken into account. If this ratio is small or $\gamma_{j|\theta}$ is no longer significant, income inequality is accounted for by the variance between the groups defined by the covariate x_j rather than by the within-groups variance. A large ratio or a significant $\gamma_{j|\theta}$ suggests that differentials in income inequality may increase if there are changes in mean income in certain subgroups.

From a public policy perspective, a small $\Delta_{\gamma_j|\theta}$ suggests that changes in the distribution of xj (e.g., increasing mean years of schooling in a population, or reducing the participation of the informal sector in the economy) may result in a more even income distribution. On the contrary, a large ratio suggests that policies that change the distribution of x_j might have small effects on income distribution, and further research is needed to understand income inequality within certain subgroups.

4. Data

We select three countries for the analysis: Uruguay (the country with the most equal distribution in the region), Brazil (a country with a high inequality), and Colombia (mid-range income inequality). We use the datasets standardized by the Luxembourg Income Study (LIS). An important advantage in using the data prepared by LIS is that the variables have been standardized for improving comparability. The original dataset from Brazil comes from Pesquisa Nacional por Amostra de Domicilios (PNAD)- 2011 (IBGE, 2011). The total sample size is 358,919, and the effective sample size for the model is 146,662 persons. The Uruguayan dataset is Encuesta Continua de Hogares ECH-2004 (INE, 2004), with a total sample size of 55,587 persons, and an effective sample size of 21,658 inhabitants. The dataset from Colombia refers to Gran Encuesta Integrada de Hogares GEIH-2010 (DANE, 2010); it has a total and an effective sample sizes of 67,190 and 24,036 persons, respectively.

The main dependent variable is labor income. Given that the analysis models log-income, we exclude all persons with zero or negative income. We selected a small set of covariates to predict income. Two independent variables are selected following the traditional Mincer model (Mincer, 1974): years of education and age. Rather than operationalizing age with age and age-square, we created three age groups: less than 30 years old, 30 to 49 years old, and 50 years old or older. This operationalization allows an easier analysis of between-groups and within-groups variance. The other selected covariates –sex, economic sector (agriculture, industry, services), occupational position (dependent employee, employer, and self-employed or own-account), and occupation (managers and professionals, other skilled workers, and laborers or elementary occupations)– have been used in other income equations for Latin American countries (Martínez-Jasso and Acevedo-Flores, 2004; Mejía, 2011). All of the covariates are used in both the mean and the variance equations. Additionally, weekly hours worked is included as a control variable in the mean equation (but its coefficient is not presented).

5. Results

The Variance-Equation-only model shows differences in income inequality between groups (as mentioned before the Variance-Equation-only model is controlling for weekly hours worked in the mean equation). The direction of the differences would be equivalent to the differences in Gini equation models. As an example, in the three countries, income inequality is greater among women than among men, and the difference is greater in Colombia than in Brazil or in Uruguay, after controlling for variance differences due to the other covariates. Table 1 also shows that income inequality is greater among older workers; among employers and self-employed work force; in the primary sector (agriculture); and among managers and professionals. Additionally, the log-income variance decreases as the years of schooling grows. In the cross-country comparison, among the few differences worth highlighting are lower inequality in the industry sector in Brazil and Uruguay, but not in Colombia; and the finding that, when compared to non-skilled laborers, mid-skilled workers have a more equal income distribution in Uruguay, a more unequal distribution in Colombia, and a similar level of income inequality in Brazil.

As is commonly known, any variance can be expressed as a sum of within-groups variance and between-groups variance. In the Variance-Equation-only model, there is no between-groups variability (except the differences by weekly hours). When the Mean-and-Variance-Equations model is estimated, the Variance Equation is modeling residual variance. The $\Delta_{\gamma_j|\theta}$ ratio measures how the differences in within group variability changes when the mean model is estimated (that is, the between-groups variance is taken into account). Ratios under one expresses that the differences among variances diminish after controlling for mean differences. In the three countries, for most of the variables the ratios are under one suggesting that income inequality may decrease if the income of poorer groups are uniformly increased, on average. Nonetheless, there are some ratios with values over one. In Brazil and Uruguay, the ratio is sizably large for females; this finding suggests that policies that increase mean income of poorer groups defined by the other covariates (age, education, occupational variables) may ameliorate income distribution for men much more than for women, thus increasing the differences in inequality between males and females. A similar pattern is found for independent workers (self-employed and employers); these groups have a more unequal income distribution than employees. Distributional changes based on the other covariates may accentuate the differences in income inequality between dependent and independent workers.

Additionally, there are some ratios that are negative. These negative ratios represent changes in the direction of the association between the covariate and log-income variance. For example, in Colombia, younger people have lower income inequality; however, after controlling for mean differences, younger people may have actually a worse income distribution than older workers. The most extreme negative ratios are observed for "other skilled workers" in Brazil, and for the industrial work force in Colombia. These finding suggests that the similarities in income distributions between the industrial and the service sectors in Colombia diverge when mean income levels are taken into account. On the contrary, Brazilian mid-skilled workers appear to have the most equal distribution after controlling for mean differences due to other covariates. From a policy perspective, the $\Delta_{\gamma_j|\theta}$ ratio suggest that economic measures to improve income distribution by uniformly increasing mean income levels in certain groups should focus on the variables with over-one ratios because if mean income is increased in other groups, income inequality might persist due to greater differential variability in the groups with large ratios.

6. Conclusions

HGLM are useful models for analyzing income inequality, because of its statistical properties in estimating a separate equation for the log-income variance as a function of covariates. From a policy perspective, the $\Delta_{\gamma_j|\theta}$ ratio suggest that economic measures to improve income distribution by uniformly increasing mean income levels in certain groups should focus on the variables with over-

one ratios because if mean income is increased in other groups, income inequality might persist due to greater differential variability in the groups with large ratios.

Table 1. Exponentiated coefficients of Variance-Equation-only model for Brazil, Colombia, and Uruguay

Covariates	Coefficients		
	Brazil	Colombia	Uruguay
Education	0.93 *	0.91 *	0.99 *
Age:			
-Younger than 30	0.93 *	0.91 *	0.99
-30 to 49 (Ref)	1.00	1.00	1.00
-50 or older	1.12 *	1.20 *	1.10 *
Females (Ref: Males)	1.02 *	1.20 *	1.04 *
Occupational position:			
Employee (Ref)	1.00	1.00	1.00
Employer	1.30 *	1.36 *	1.41 *
Self-employed	1.30 *	1.21 *	1.26 *
Economic Sector:			
-Agriculture	1.33 *	1.12 *	1.19 *
-Industry	0.87 *	1.00	0.97 *
-Services (Ref)	1.00	1.00	1.00
Occupation			
-Managers and professionals	1.77 *	1.42 *	1.13 *
-Other skilled workers	0.99	1.05 *	0.88 *
-Laborers (Ref)	1.00	1.00	1.00
Constant	0.65 *	0.48 *	0.66 *

Source: Luxembourg Income Study LIS
Notes: *: $p<0.05$

Table 2. Ratio of Variance Equation coefficients controlling for conditional means, divided by coefficients in Intercept-only equation, for Brazil, Colombia, and Uruguay

Covariates	$\Delta_{\gamma_j\|\theta}$ ratio		
	Brazil	Colombia	Uruguay
Education	-0.54	0.19	0.08
Age:			
-Younger than 30	0.29	-0.57	-0.33
-30 to 49 (Ref)			
-50 or older	0.81	0.55	0.80
Females (Ref: Males)	4.27	0.78	1.49
Occupational position:			
Employee (Ref)			
Employer	1.16	1.15	0.65
Self-employed	1.24	1.85	1.31
Economic Sector:			
-Agriculture	0.40	0.33	0.50
-Industry	0.84	-2.83	1.02
-Services (Ref)			
Occupation			
-Managers and professionals	0.65	0.25	0.05
-Other skilled workers	-25.83	0.73	0.12
-Laborers (Ref)	-0.54	0.19	0.08

Source: Luxembourg Income Study LIS

References

Allison, P. D. (1978). Measures of inequality. *American Sociological Review*, 43(6), 865-880.

Departamento Administrativo Nacional de Estadística DANE (2010). Great Integrated Household Survey /Gran Encuesta Integrada de Hogares-GEIH 2010.

Instituto Brasileiro de Geografia e Estatística IBGE (2011). National Household Sample Survey/Pesquisa Nacional por Amostra de Domicílios - PNAD 2011

Instituto Nacional de Estadística - INE (2004). Continuous Household Survey/Encuesta Continua de Hogares – ECH 2004.

Martínez Jasso, I., & Acevedo Flores, G. J. (2004). La brecha salarial en México con enfoque de género: capital humano, discriminación y selección muestral. *Ciencia UANL*, *7*(1), 66-71.

Mejía, L. B. (2011). Diferencias regionales en la distribución del ingreso en Colombia. *Revista Sociedad y Economía*, (21), 43-68.

Mincer,J.(1974). *Schooling, Experience and Earnings*. New York: National Bureau of Economic Research.

Nelder, J.A., Wedderburn, R.W.M. (1972). "Generalized Linear Models". *Journal of the Royal Statistical Society. Series A (General)*, 135(3): 370-384.

Smyth, G. K. (1989). Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47-60.

Western, B., & Bloome, D. (2009). Variance function regressions for studying inequality. *Sociological Methodology*, 39(1), 293-326.

Zheng H, Yang Y, Land KC (2013). "Heteroscedastic regression models for the systematic analysis of residual variances". In: Morgan, S. *Handbook of Causal Analysis for Social Research.* New York, NY: Springer, Pp. 133-152.