the review of income and wealth

Review of Income and Wealth
Series 59, Number 3, September 2013
DOI: 10.1111/roiw.12045

INEQUALITY OF OPPORTUNITY IN BRAZIL: A CORRIGENDUM

BY FRANÇOIS BOURGUIGNON

Paris School of Economics

FRANCISCO H. G. FERREIRA*

The World Bank and IZA

AND

MARTA MENÉNDEZ

Université Paris Dauphine and DIAL

This note acknowledges and corrects a programming error in our paper "Inequality of Opportunity in Brazil" (*Review of Income and Wealth*, 53(4), 585–618, 2007). Once the error is corrected, our bounds approach to the identification of individual model parameters in the presence of omitted variable biases is much less useful than indicated in the original paper. In the specific context of the measurement of inequality of opportunity, this implies that the decomposition of overall inequality of opportunity into direct and indirect effects is not reliable. However, the parametric approach introduced in our paper remains useful for obtaining a lower-bound estimate of *overall* ex-ante inequality of opportunity, as proposed by Ferreira and Gignoux (2011).

JEL Codes: D31, D63

Keywords: Brazil, inequality of opportunities, omitted variable bias

Our paper "Inequality of Opportunity in Brazil" (*Review of Income and Wealth*, 53(4), 585–618, 2007) contains a non-trivial error.[1] In that paper, we proposed a measure of inequality of opportunity as the share of earnings (w) inequality explained by predetermined, morally irrelevant circumstances (C). The main results of the paper were obtained from the OLS estimation of a reduced-form model given by:

(equation 10 in the paper) $$\ln w_i = C_i\psi + \varepsilon_i.$$

We denoted a counterfactual earnings distribution where all differences in circumstances were eliminated as $\tilde{\Phi}(\tilde{w})$, with $\tilde{w}_i = \exp[\bar{C}\hat{\psi} + \hat{\varepsilon}_i]$.[2] If the actual earnings distribution is given by $\Phi(w)$, we proposed to measure inequality of opportunity in that distribution by the ratio $\Theta_I := \dfrac{I(\Phi) - I(\tilde{\Phi})}{I(\Phi)}$, where I denotes some well-behaved inequality measure, such as the Theil index. This is an indirect

*Correspondence to: Francisco Ferreira, Development Research Group, The World Bank, 1818 H Street, NW, Washington DC 20433, USA (fferreira@worldbank.org).

[1]We are very grateful to Esteban Puentes for first pointing the error out to us.

[2]A hat denotes an OLS estimate and an overbar denotes an arithmetic mean.

approach: $I(\tilde{\Phi})$ captures the inequality that remains when all inequality of opportunity (i.e., between people with different circumstances) is eliminated. So $I(\Phi) - I(\tilde{\Phi})$, or the ratio of that difference to the total, are measures of inequality of opportunity.

Because equation (10) was the reduced form of a model containing effort as well as circumstance variables, this measure of inequality of opportunity should reflect both the direct effects of circumstances on earnings, and the indirect effects operating through efforts (*E*). To distinguish between those two categories of effects, we also estimated:

(equation 5′) $$\ln w_i = C_i\alpha + E_i\beta + u_i.$$

We recognized that the existence of omitted circumstance variables would bias the OLS estimates of ψ, and that omitted circumstance and effort variables would bias the estimates of α and β. We argued that suitable instruments were not available and proposed instead to investigate the likely magnitude of potential biases, by estimating upper and lower bounds both for the true coefficients and for the measures of inequality of opportunity, which were the main object of interest.

Focusing on equation (10) in the original paper, if the error term ε is not orthogonal to C (but the two are jointly normally distributed), then the estimated vector of coefficients $\hat{\psi}$ is biased, and the bias can be written as:

$$B = E(\hat{\psi}) - \psi = \sum\nolimits_C^{-1}\left(\underline{\rho_{C\varepsilon}\sigma_c}\right)\sigma_\varepsilon$$

where Σ_X denotes the theoretical variance–covariance matrix of a random vector X, σ_x denotes the standard deviation of a variable x, and ρ_{xy} denotes the theoretical correlation coefficient between two variables x and y or the vector of correlation coefficients between a vector x and a variable y. Because these theoretical population parameters are unknown, our proposed solution was to evaluate the approximate size of the bias by:

$$B \cong N(C'C)^{-1}\left(\underline{\tilde{\rho}_{C\varepsilon}\hat{\sigma}_c}\right)\tilde{\sigma}_\varepsilon.$$

To compute this sample-based approximation, we calculated: $\tilde{\sigma}_\varepsilon^2 = \hat{\sigma}_\varepsilon^2/(1-K)$, where $\hat{\sigma}_\varepsilon^2$ is the variance of the OLS residual of the regressions above and $K = (\underline{\tilde{\rho}_{C\varepsilon}\hat{\sigma}_c})'(C'C)^{-1}(\underline{\tilde{\rho}_{C\varepsilon}\hat{\sigma}_c})$. $\tilde{\rho}_{C\varepsilon}$ denotes drawings from a uniform distribution defined on (–1, 1), with any values such that $K \geq 1$ being rejected. Finally, we also imposed a set of additional constraints on the signs of coefficient estimates (empirically backed by the literature). Please see the original paper for a more detailed description of the method. Using this approach, we reported bounds around both the regression coefficients and the measures of inequality of opportunity which (we hoped) were sufficiently narrow as to be informative.

Unfortunately, our calculation of the range of possible values for the biases in both equations (10) and (5′) contained a mistake. When empirically estimating $\tilde{\sigma}_\varepsilon^2 = \hat{\sigma}_\varepsilon^2/(1-K)$, a programming misspelling we made in Stata led us to use the standard error of the linear prediction (command option "stdp"), instead the standard error of the residual (command option "stdr"). This programming error

led us to underestimate the value of $\hat{\sigma}_{\varepsilon}^{2}$ by a factor ranging from 37 to 92 (depending on the cohort considered).

When the error is corrected and the biases are recomputed, the bounds around the OLS estimates of the regression coefficients become much wider. The small set of conditions we had previously imposed on coefficients now proves insufficient to obtain informative bounds. An alternative approach, which illustrates how the "confidence intervals" widen as we move away from OLS assumptions, is to draw the correlation coefficients $\tilde{\rho}_{C\varepsilon}$ for all circumstance variables from uniform distributions defined sequentially on broader supports: (–0.05, 0.05), (–0.1, 0.1), (–0.15, 0.15), and (–0.2, 0.2). Note, however, that these supports are all much narrower than the widest possible range used earlier: (–1, 1). Results from this approach are presented in Table 1 for selected regression coefficients (those on mean parental schooling), and in Table 2 for $I(\tilde{\Phi})$, our measure of counterfactual inequality when all inequality due to circumstances is eliminated.

Two implications arise from this exercise. First, once our coding error is corrected, the bounds approach employed in our original paper no longer appears useful for identifying a narrow range of possible values for the biases plaguing OLS regression coefficients. There is no rationale for restricting the possible correlation between explanatory variables and a regression residual ex-ante to a narrow interval such as (–0.2, 0.2). The true value of $\rho_{C\varepsilon} \in (-1,1)$ is, of course, unknown. When we allow for the full possible range of values for that correlation coefficient, our use of sample moments to calculate approximate bounds on the value of the bias of OLS coefficients turns out to yield intervals that are too large to be of any practical use.

Second, the effect of correcting the error on the bounds around the estimates of counterfactual inequality—and in particular on the lower-bound estimate—is much less pronounced. In fact, as shown in Table 2, the lower-bound on the Theil coefficient of inequality when differences in circumstances are eliminated is quite robust to changes in the assumed correlation coefficients between circumstance variables and the regression residual.

The Krishnakumar Correction

After Mr. Esteban Puentes kindly pointed out our programming error to us, but before we had finished this corrigendum, we became aware of a note proposing a "correction" of our 2007 paper (Krishnakumar, 2013). That note, which is being published alongside this corrigendum, makes a number of notational corrections, which we largely accept. We should indeed have made the assumption of joint normality of C and ε explicit (or used probability limits and referred to the asymptotic bias), and used clearer notation to distinguish between population and sample moments.

However, contrary to what the note suggests, notational imprecision was *not* responsible for the error in our paper. In particular, we never estimated or reported what Krishnakumar (2013) calls "the BFM bias" in her Table 1. From the outset, our estimates of the bias were what she calls "the corrected BFM bias" for which, as she notes: ". . . for a known ρ_{xu}, the theoretical bias, the small sample bias and the corrected BFM bias (with the 1/N factor) are all of the same order of

TABLE 1

COEFFICIENTS OF MEAN PARENTAL YEARS OF SCHOOLING BY COHORT, REDUCED-FORM MODEL[a,b]

Mean parental schooling (years)	b1936_40	b1941_45	b1946_50	b1951_55	b1956_60	b1961_65	b1966_70
Upper bound estimates							
$-0.2 \leq$ rho (Xi, u) ≤ 0.2	0.265	0.195	0.198	0.185	0.163	0.149	0.136
$-0.15 \leq$ rho (Xi, u) ≤ 0.15	0.242	0.186	0.188	0.170	0.151	0.140	0.126
$-0.1 \leq$ rho (Xi, u) ≤ 0.1	0.218	0.174	0.174	0.154	0.137	0.127	0.113
$-0.05 \leq$ rho (Xi, u) ≤ 0.05	0.195	0.157	0.157	0.138	0.123	0.114	0.102
OLS estimates	**0.162*****	**0.137*****	**0.143*****	**0.119*****	**0.103*****	**0.103*****	**0.088*****
Lower bound estimates							
$-0.05 \leq$ rho (Xi, u) ≤ 0.05	0.135	0.109	0.108	0.093	0.082	0.077	0.067
$-0.1 \leq$ rho (Xi, u) ≤ 0.1	0.097	0.077	0.075	0.062	0.054	0.052	0.043
$-0.15 \leq$ rho (Xi, u) ≤ 0.15	0.057	0.043	0.039	0.028	0.025	0.025	0.018
$-0.2 \leq$ rho (Xi, u) ≤ 0.2	0.012	0.005	−0.001	−0.009	−0.009	−0.006	−0.011

[a]Our dependent variable is the log of hourly wage rate, and explanatory variables include race, parental schooling (mean and difference from mother's and father's), regional dummies and father's occupational status. [b]For our selected variable, mean parental years of schooling , we present the following values: the minimum and maximum coefficient estimates from the 90th confidence intervals of simulations, using four possible value intervals for correlation coefficients of our X's and the residuals: (−0.05, +0.05), (−0.1, +0.1), (−0.15, +0.15), (−0.2, +0.2); the OLS estimates and significance levels is in between; *significant at 10%; **significant at 5%; ***significant at 1%.

TABLE 2

EARNINGS INEQUALITY WHEN INEQUALITY OF OPPORTUNITY IS ELIMINATED, URBAN MEN IN BRAZIL: COUNTERFACTUAL THEIL COEFFICIENTS

	b1936_40	b1941_45	b1946_50	b1951_55	b1956_60	b1961_65	b1966_70
Total Observed Inequality	**0.873**	**0.997**	**0.759**	**0.655**	**0.706**	**0.580**	**0.566**
Counterfactual inequality when circumstances are equalized							
Upper bound estimates							
$-0.2 \leq$ rho (Xi, u) ≤ 0.2	0.754	0.778	0.710	0.602	0.658	0.592	0.592
$-0.15 \leq$ rho (Xi, u) ≤ 0.15	0.717	0.734	0.672	0.561	0.619	0.442	0.553
$-0.1 \leq$ rho (Xi, u) ≤ 0.1	0.688	0.698	0.645	0.537	0.595	0.421	0.526
$-0.05 \leq$ rho (Xi, u) ≤ 0.05	0.667	0.673	0.627	0.525	0.575	0.410	0.507
Mean estimates	**0.654**	**0.656**	**0.619**	**0.519**	**0.562**	**0.407**	**0.494**
Lower bound estimates							
$-0.05 \leq$ rho (Xi, u) ≤ 0.05	0.647	0.641	0.606	0.518	0.558	0.403	0.489
$-0.1 \leq$ rho (Xi, u) ≤ 0.1	0.638	0.633	0.602	0.521	0.558	0.405	0.486
$-0.15 \leq$ rho (Xi, u) ≤ 0.15	0.638	0.629	0.601	0.526	0.561	0.412	0.485
$-0.2 \leq$ rho (Xi, u) ≤ 0.2	0.645	0.629	0.620	0.533	0.567	0.428	0.487

magnitude." Neither is it the case that our bounds approach would have yielded complex bounds, as suggested in her Tables 2–4. The author ignores a crucial step in our approach, which was to discard any drawings of $\tilde{\rho}_{C\varepsilon}$ for which $K \geq 1$.

The error was not due to any of the points 1–4 in the note. It is due to the unfortunate Stata coding error described above. Whatever the reason, however, Krishnakumar (2013) is right in her final claim that ". . . the confidence intervals presented in Bourguignon, Ferreira and Menéndez (2007) is not correct, and the results do not provide the correct range of bias of their OLS estimates."

Implications for the Measurement of Inequality of Opportunity

This unfortunate error, for which we apologize to our readers, implies that our bounds approach to the identification of individual model parameters in the presence of omitted variable biases is much less useful than indicated in the original paper. In the specific context of the measurement of inequality of opportunity, this means that the decomposition of overall inequality of opportunity into direct and indirect effects (as in Panel 2 of Table 5) is not reliable. Neither are estimates of the contribution of individual circumstance variables to earnings inequality (Table 6).

The error does *not* imply, however, that this parametric approach to measuring *overall* ex-ante inequality of opportunity is useless. In a subsequent paper, heavily inspired by our 2007 paper, Ferreira and Gignoux (2011) have proposed using inequality in the *predicted* incomes from equation (10) as a direct measure of inequality of opportunity: $I(\exp[C_i\hat{\psi}])$. Those authors refer to this level measure of inequality of opportunity as IOL, and to its ratio to total observed inequality as IOR. Ferreira and Gignoux (2011) acknowledge that sub-decompositions of these measures into direct or indirect effects, or into the effects of individual circumstances, would require strong assumptions about the orthogonality of residuals in (10). But they also show that IOL and IOR can safely be interpreted as lower-bound estimates of *overall* inequality of opportunity—i.e., inequality due to *all* predetermined circumstances, not only to those that are observed. A formal proof is provided. For a more recent attempt at disentangling direct and indirect effects of circumstances on final outcomes, subject to its own set of assumptions, see Björklund *et al*. (2011).

References

Björklund, A., M. Jäntti, and J. Roemer, "Equality of Opportunity and Distribution of Long-Run Income in Sweden," IZA Discussion Paper No. 5466, 2011.

Bourguignon, F., F. H. G. Ferreira, and M. Menéndez, "Inequality of Opportunity in Brazil," *Review of Income and Wealth*, 53(4), 585–618, 2007.

Ferreira, F. H. G. and J. Gignoux, "The Measurement of Inequality of Opportunity: Theory and an Application to Latin America," *Review of Income and Wealth*, 57(4), 622–57, 2011.

Krishnakumar, J., "On Endogeneity Bias: Some General Remarks and Correction of the ROIW Paper by Bourguignon, Ferreira and Menéndez (2007)," *Review of Income and Wealth*, 59(3), 2013.