sociological science

Supplement to:

Senkevics, Adriano S., Rogério J. Barbosa, Flavio Carvalhaes, and Carlos A. Costa Ribeiro. 2024. "Decomposing Heterogeneity in Inequality of Educational Opportunities: Family Income and Academic Performance in Brazilian Higher Education." Sociological Science 11: 854-885.

Statistical Appendix

A – Raking (Iterative Proportional Fitting)

Iterative proportional fitting raking is a statistical technique usually employed in sampling to adjust weights. Its primary objective is to align the weighted (sample) data with known (population) margins across multiple dimensions. In our case, we want to recalibrate weights making them match the performance distribution conditional on income quantiles of interest, while keeping the marginals of all other variables unchanged. The algorithm can be described iteratively as follows:

1. **Preliminary: Obtaining a list of Target Margins**
   - 1.1. Using the whole sample and the initial weights w_i, calculate the observed weighted marginal frequency distribution of all variables in the dataset that were used as regressors in the multinomial model (except Performance and Income).
   - 1.2. Calculate the weighted frequency distribution of Performance, conditional on $Income = U$.
2. **Raking Adjustment for variable X_k:**
   - 2.1. Take the dataset with its weights w_i
   - 2.2. Pick variable X_k from the list of Target Margins. Assign its marginal frequencies to the individuals in the source dataset, matching the categories they belong to. This new variable contains the Target Margins of variable X_k.
   - 2.3. Calculate the weighted frequency distribution of Performance using the weights current available in the data set and also assign these values to the individuals matching the categories they belong to. This new variable contains the Source Margins of variable X_k.
   - 2.4. Make:

$$w_i^{P \vee \text{Income}=U} = w_i \times \frac{\left(\text{Target Margins of variable } X_k\right)}{\left(\text{Source Margins of variable } X_k\right)}$$

   - 2.5. Repeat until convergence.

B – Inference and Confidence intervals

A relevant question is how statistically significant are the *INE* results. This can be easily pursued via bootstrap: one can resample from the full dataset and run S times the computation of the quantities of interest. This approach has the advantage of sources of variation of both the predicted probability of making a choice ($P_{U,d}$ and $P_{L,d}$) as well as the performance composition within income strata ($I_{U,d}$ and $I_{L,d}$). However this implies running the model and data analysis multiple times, which is computationally demanding, considering our big sample size.

An alternative is to regard $I_{U,d}$ and $I_{L,d}$ as fixed and proceed a Parametric Bootstrap for estimating bounds for $P_{U,d}$ and $P_{L,d}$, which are average predicted probabilities from a multinomial regression. As the sample is comprised by millions of individual, the sample error of performance composition is very small – then, in practice, assuming it is fixed does not imply a underestimation of the pursued sample errors.

In a 3-choice model with $K-1$ predictors and an intercept, there is $2K$ estimated coefficients:

$$\underbrace{\hat{\beta}}_{2K\times1} = \begin{bmatrix} \underbrace{\hat{\beta}_{private}}_{K\times1} \\ \text{¿}\underbrace{\hat{\beta}_{public}}_{K\times1} \end{bmatrix}$$

From the Central Limit Theorem, we know that $\hat{\beta} \;\; N\left(\beta, \underbrace{\Sigma}_{2K\times2K}\right)$, where Σ is the variance-covariance matrix of the coefficients. Using $\hat{\beta}$ as the mean vector of a multidimensional Normal Distribution and $\hat{\Sigma}$ as the variance-covariance matrix, we drew 2000 samples (S) of coefficients with a multivariate normal random number generator, and gathered them into the $\underbrace{\hat{B}^{S}_{private}}_{K\times S}$ and $\underbrace{\hat{B}^{S}_{public}}_{K\times S}$ matrices. These coefficients produced predicted probabilities 000 different estimates for each individual in the data set, which were then used to produce a distribution of the *INE* index and the other quantities of interest.

Annex

Table A1. Descriptive statistics of young high-school graduates, cohort 2012

	All the cohort	*No Access*	*Public*	*Private*
N	1,133,027	351,715	188,863	592,449
%	100.0	31.0	16.7	52.3
Educational performance				
Mean	506.8	462.1	580.4	509.9
Standard deviation	77.6	55.9	80.7	68.3
Income deciles (%)				
1	10.0	18.8	5.8	6.8
2	10.0	14.7	7.9	7.8
3	10.0	13.2	9.1	7.0
4	10.0	14.0	8.6	7.1
5	10.0	9.3	10.9	8.5
6	10.0	9,9	10.6	8.3
7	10.0	8.3	11.3	9.0
8	10.0	6.2	11.9	11.0
9	10.0	4.0	12.2	14.2
10	10.0	1.6	11.7	20.3
Parental education (%)				
Less than high school	41.6	63.9	24.9	33.7
Completed high school	35.8	29.3	35.6	39.7
Higher education or more	22.6	6.8	39.5	26.6
Sex (%)				
Female	59.2	59.0	54.4	60.8
Male	40.8	41.0	45.6	39.2
Age				
Mean	17.6	17.9	17.3	17.5
Standard-deviation	1.0	1.2	0.8	0.8
Rural household (%)				
No	97.4	96.0	97.7	98.2
Yes	2.6	4.0	2.3	1.8
Type of high-school (%)				
Public, state	76.4	94.5	53.5	72.9
Public, city	0.9	0.8	0.8	1.0
Public, federal	1.2	0.4	4.5	0.7
Private	21.5	4.3	41.2	25.4

Table A2. Pairwise comparisons between all upper (rows) and lower income strata (columns): INE, Primary Effect, and Compensatory Advantage

Upper Income Decile	INE									Primary Effect									Compensatory Advantages								
	Lower Income Decile									*Lower Income Decile*									*Lower Income Decile*								
(Reference)	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
No Access																											
2	-0,09	-	-	-	-	-	-	-	-	-0,04	-	-	-	-	-	-	-	-	-0,04	-	-	-	-	-	-	-	-
3	-0,18	-0,09	-	-	-	-	-	-	-	-0,08	-0,04	-	-	-	-	-	-	-	-0,10	-0,06	-	-	-	-	-	-	-
4	-0,16	-0,06	0,02	-	-	-	-	-	-	-0,08	-0,03	0,01	-	-	-	-	-	-	-0,08	-0,04	0,01	-	-	-	-	-	-
5	-0,23	-0,14	-0,06	-0,08	-	-	-	-	-	-0,11	-0,07	-0,03	-0,04	-	-	-	-	-	-0,11	-0,07	-0,02	-0,03	-	-	-	-	-
6	-0,23	-0,14	-0,05	-0,07	0,00	-	-	-	-	-0,12	-0,08	-0,04	-0,04	-0,01	-	-	-	-	-0,11	-0,07	-0,02	-0,03	0,00	-	-	-	-
7	-0,28	-0,19	-0,10	-0,12	-0,05	-0,05	-	-	-	-0,15	-0,10	-0,06	-0,07	-0,03	-0,03	-	-	-	-0,14	-0,10	-0,05	-0,06	-0,03	-0,03	-	-	-
8	-0,31	-0,22	-0,13	-0,15	-0,08	-0,08	-0,03	-	-	-0,17	-0,13	-0,09	-0,09	-0,05	-0,05	-0,02	-	-	-0,15	-0,11	-0,06	-0,07	-0,04	-0,04	-0,01	-	-
9	-0,36	-0,27	-0,18	-0,20	-0,13	-0,13	-0,08	-0,05	-	-0,20	-0,16	-0,12	-0,12	-0,09	-0,08	-0,05	-0,03	-	-0,17	-0,14	-0,09	-0,09	-0,07	-0,07	-0,04	-0,03	-
10	-0,42	-0,33	-0,24	-0,26	-0,19	-0,19	-0,14	-0,11	-0,06	-0,27	-0,22	-0,17	-0,18	-0,14	-0,14	-0,11	-0,08	-0,05	-0,18	-0,14	-0,10	-0,10	-0,08	-0,08	-0,05	-0,04	-0,02
Public																											
2	0,02	-	-	-	-	-	-	-	-	0,02	-	-	-	-	-	-	-	-	0,00	-	-	-	-	-	-	-	-
3	0,02	0,01	-	-	-	-	-	-	-	0,04	0,02	-	-	-	-	-	-	-	-0,01	-0,01	-	-	-	-	-	-	-
4	0,03	0,01	0,00	-	-	-	-	-	-	0,04	0,02	0,00	-	-	-	-	-	-	0,01	0,01	0,02	-	-	-	-	-	-
5	0,05	0,03	0,02	0,02	-	-	-	-	-	0,07	0,05	0,03	0,03	-	-	-	-	-	-0,02	-0,02	-0,01	-0,03	-	-	-	-	-
6	0,05	0,03	0,03	0,02	0,00	-	-	-	-	0,07	0,05	0,03	0,03	0,01	-	-	-	-	-0,03	-0,03	-0,02	-0,04	-0,01	-	-	-	-
7	0,07	0,06	0,05	0,05	0,03	0,02	-	-	-	0,10	0,08	0,06	0,06	0,03	0,02	-	-	-	-0,05	-0,05	-0,04	-0,05	-0,02	-0,01	-	-	-
8	0,10	0,09	0,08	0,08	0,06	0,05	0,03	-	-	0,13	0,11	0,09	0,09	0,06	0,05	0,03	-	-	-0,06	-0,06	-0,05	-0,07	-0,04	-0,03	-0,01	-	-
9	0,16	0,14	0,13	0,13	0,11	0,11	0,09	0,06	-	0,18	0,16	0,14	0,14	0,11	0,11	0,08	0,05	-	-0,07	-0,07	-0,06	-0,08	-0,05	-0,04	-0,03	-0,01	-
10	0,26	0,25	0,24	0,24	0,22	0,22	0,19	0,16	0,11	0,29	0,27	0,26	0,25	0,23	0,23	0,21	0,18	0,13	-0,04	-0,04	-0,04	-0,05	-0,03	-0,02	-0,01	0,00	0,01
Private																											
2	0,08	-	-	-	-	-	-	-	-	0,02	-	-	-	-	-	-	-	-	0,04	-	-	-	-	-	-	-	-
3	0,15	0,08	-	-	-	-	-	-	-	0,04	0,02	-	-	-	-	-	-	-	0,11	0,07	-	-	-	-	-	-	-
4	0,13	0,05	-0,02	-	-	-	-	-	-	0,04	0,01	-0,01	-	-	-	-	-	-	0,08	0,04	-0,03	-	-	-	-	-	-
5	0,19	0,11	0,03	0,06	-	-	-	-	-	0,05	0,03	0,01	0,01	-	-	-	-	-	0,14	0,10	0,03	0,06	-	-	-	-	-
6	0,18	0,10	0,03	0,05	-0,01	-	-	-	-	0,05	0,03	0,00	0,01	0,00	-	-	-	-	0,14	0,10	0,04	0,07	0,01	-	-	-	-
7	0,21	0,13	0,06	0,08	0,02	0,03	-	-	-	0,05	0,03	0,01	0,01	0,00	0,00	-	-	-	0,19	0,15	0,09	0,11	0,05	0,04	-	-	-
8	0,21	0,13	0,05	0,08	0,02	0,03	0,00	-	-	0,05	0,02	0,00	0,01	-0,01	0,00	-0,01	-	-	0,21	0,17	0,11	0,13	0,08	0,06	0,02	-	-
9	0,20	0,13	0,05	0,07	0,01	0,02	-0,01	-0,01	-	0,02	0,00	-0,02	-0,01	-0,03	-0,03	-0,03	-0,02	-	0,25	0,21	0,15	0,17	0,12	0,10	0,06	0,04	-
10	0,16	0,08	0,00	0,02	-0,03	-0,03	-0,05	-0,05	-0,05	-0,03	-0,05	-0,09	-0,07	-0,09	-0,09	-0,10	-0,10	-0,08	0,22	0,18	0,13	0,15	0,11	0,09	0,06	0,04	0,01

Source: Prepared by the authors based on cross-referencing databases from the CEB, Enem, and CES (Inep)