The global distribution of education: 1970-2010

Vanesa Jordá
University of Cantabria

Winter School on Inequality and Social Welfare Theory

Alba di Canazei, Italy
15th January 2015

Main contribution

- We introduce a new database for mean years of schooling and inequality measures of educational outcomes for 143 countries from 1970 to 2010.
  - **Reliability.** More accurate estimates than the existing data sets.
  - **Distribution sensitive.** A battery of GE measures for different parameter values.
  - **Flexibility.** Overall inequality can be computed for any group of countries
- We present for the first time a complete and continuous picture of the distribution of schooling.

Main contribution

- We introduce a new database for mean years of schooling and inequality measures of educational outcomes for 143 countries from 1970 to 2010.
  - **Reliability.** More accurate estimates than the existing data sets.
  - **Distribution sensitive.** A battery of GE measures for different parameter values.
  - **Flexibility.** Overall inequality can be computed for any group of countries
- We present for the first time a complete and continuous picture of the distribution of schooling.

Main contribution

- We introduce a new database for mean years of schooling and inequality measures of educational outcomes for 143 countries from 1970 to 2010.
  - **Reliability.** More accurate estimates than the existing data sets.
  - **Distribution sensitive.** A battery of GE measures for different parameter values.
  - **Flexibility.** Overall inequality can be computed for any group of countries
- We present for the first time a complete and continuous picture of the distribution of schooling.

Contents

1 Introduction

- Education is acknowledged to be a fundamental component of well-being.
  Acemoglu and angrist, 2001; Oreopoluos and Salvanes, 2011.
- MYS is a widely used indicator to measure the educational performance of a country.
  - Barro and Lee (2013).
  - Cohen and Soto (2007).
  - UNESCO (2014).

BIASED ESTIMATORS OF THE MEAN

1 Introduction

- Education is acknowledged to be a fundamental component of well-being.
  Acemoglu and angrist, 2001; Oreopoluos and Salvanes, 2011.
- MYS is a widely used indicator to measure the educational performance of a country.
  - Barro and Lee (2013).
  - Cohen and Soto (2007).
  - UNESCO (2014).

BIASED ESTIMATORS OF THE MEAN

1 Introduction

- Education is acknowledged to be a fundamental component of well-being.
  Acemoglu and angrist, 2001; Oreopoluos and Salvanes, 2011.
- MYS is a widely used indicator to measure the educational performance of a country.
  - Barro and Lee (2013).
  - Cohen and Soto (2007).
  - UNESCO (2014).

BIASED ESTIMATORS OF THE MEAN

1 Introduction

- Education is acknowledged to be a fundamental component of well-being.
  Acemoglu and angrist, 2001; Oreopoluos and Salvanes, 2011.
- MYS is a widely used indicator to measure the educational performance of a country.
  - Barro and Lee (2013).
  - Cohen and Soto (2007).
  - UNESCO (2014).

BIASED ESTIMATORS OF THE MEAN

Mean years of schooling

The main approach: weighted average.

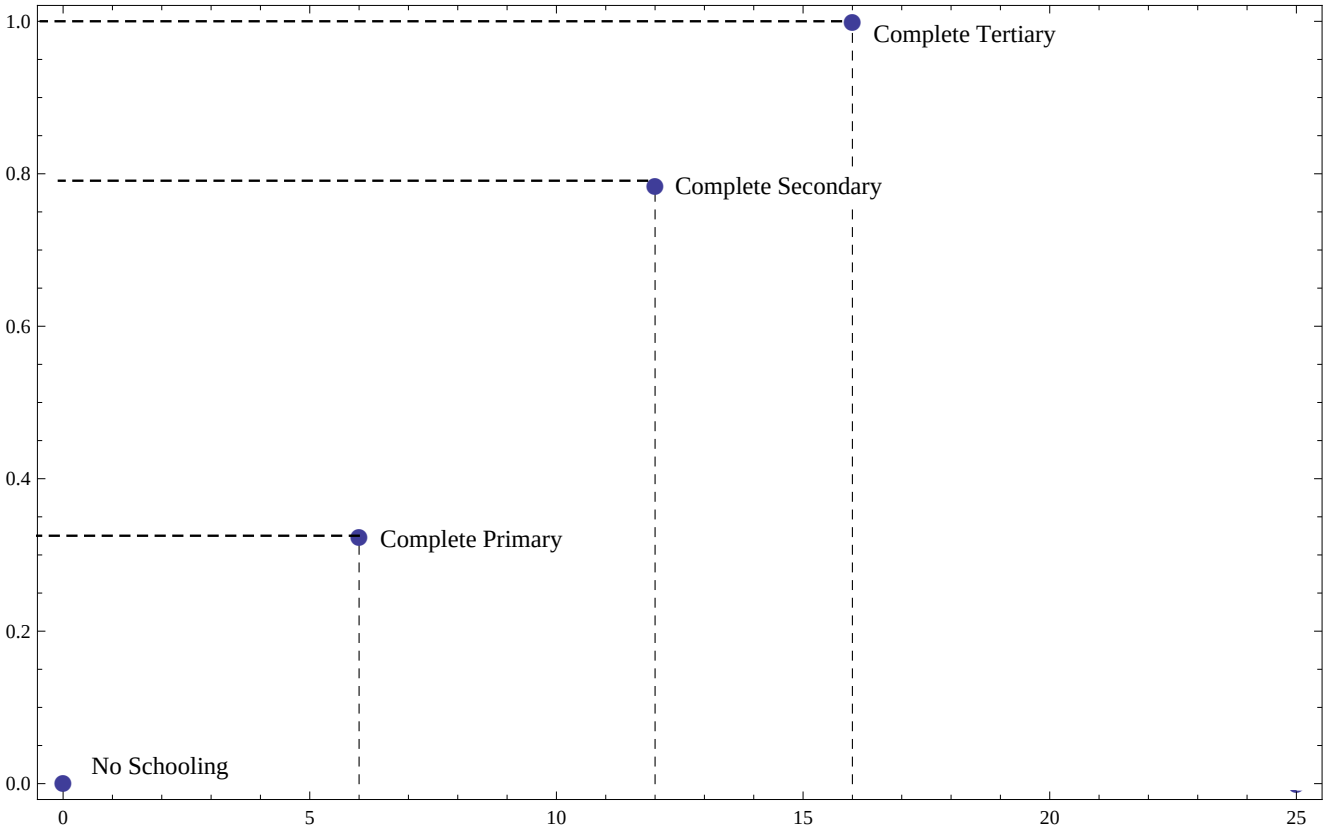
- Educational attainment rates. (census and surveys)
  - No schooling
  - Primary
  - Secondary
  - Tertiary
- Official duration of each educational level.

Example: Japan (1990)

Example: Japan (1990).

Educational level	Attainment rate	Official duration
No schooling	0.0024	0
Primary	0.3213	6
Secondary	0.4609	12
Tertiary	0.2153	16

Example: Japan (1990)



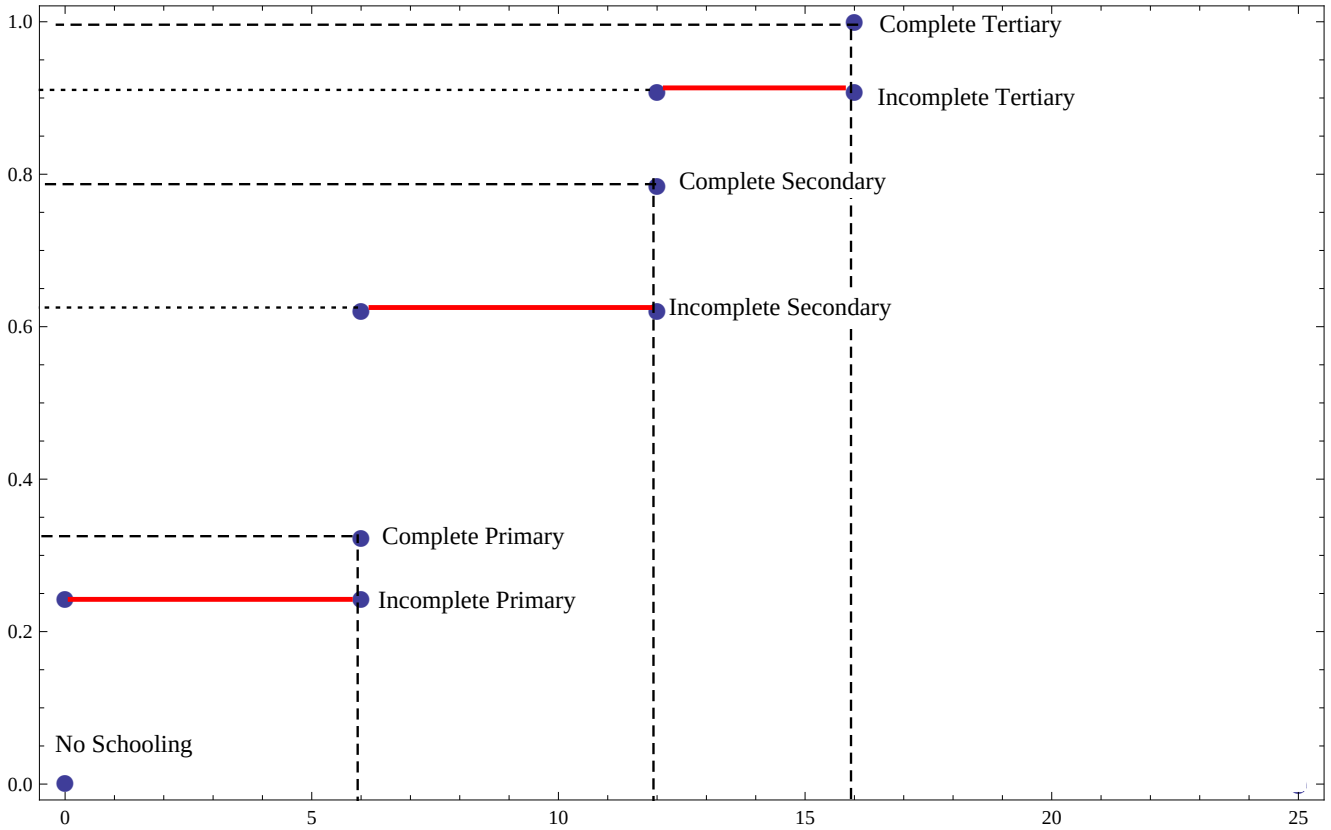
Overestimation: Not all individuals completed the educational cycle

Example: Japan (1990)

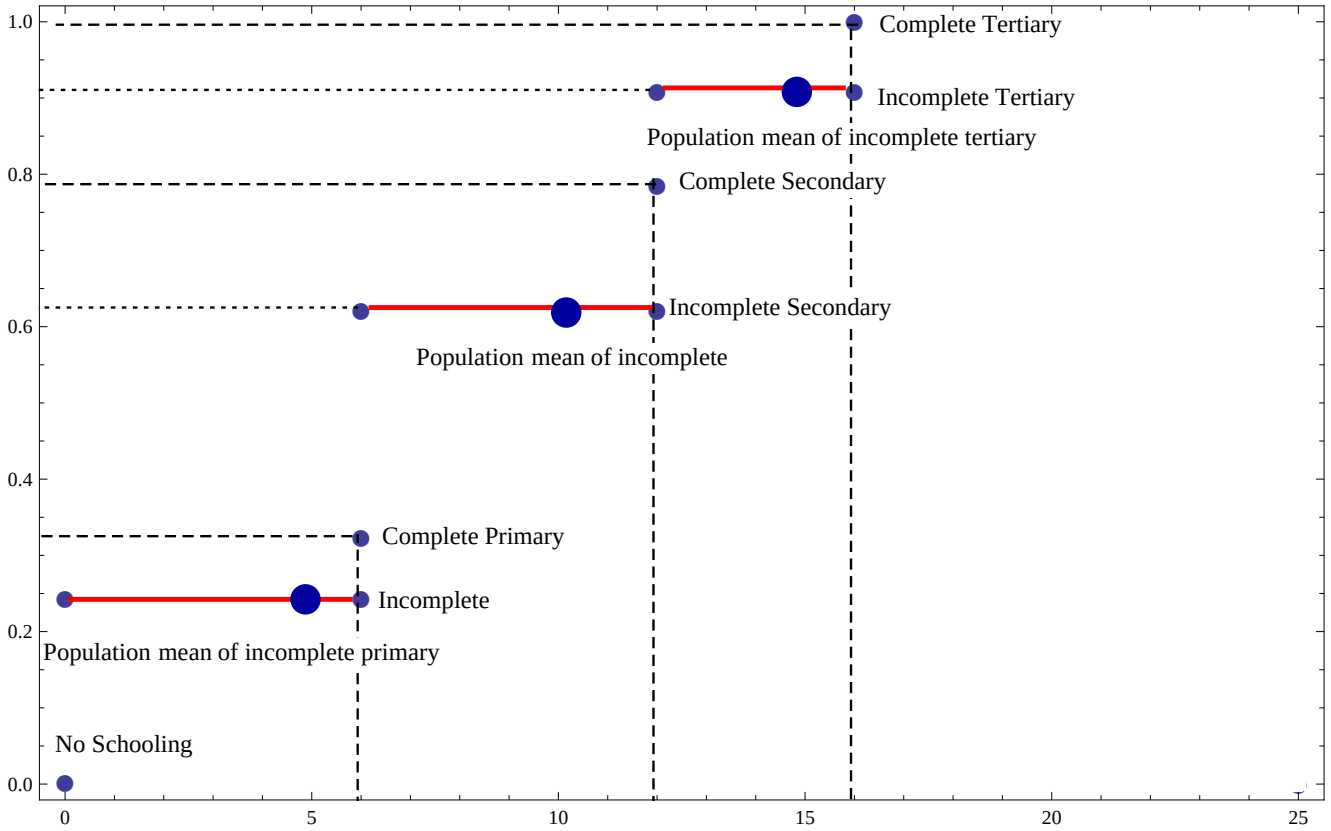
SOLUTION: Educational levels are broken down into **complete** and **incomplete** (estimated using completion rates: source of bias)

Educational level	Attainment rate	Official duration
No schooling	0.0024	0
Primary incomplete	0.2413	?
Primary complete	0.0800	6
Secondary incomplete	0.2973	?
Secondary complete	0.1636	12
Tertiary incomplete	0.1248	?
Tertiary incomplete	0.0905	16

Example: Japan (1990)

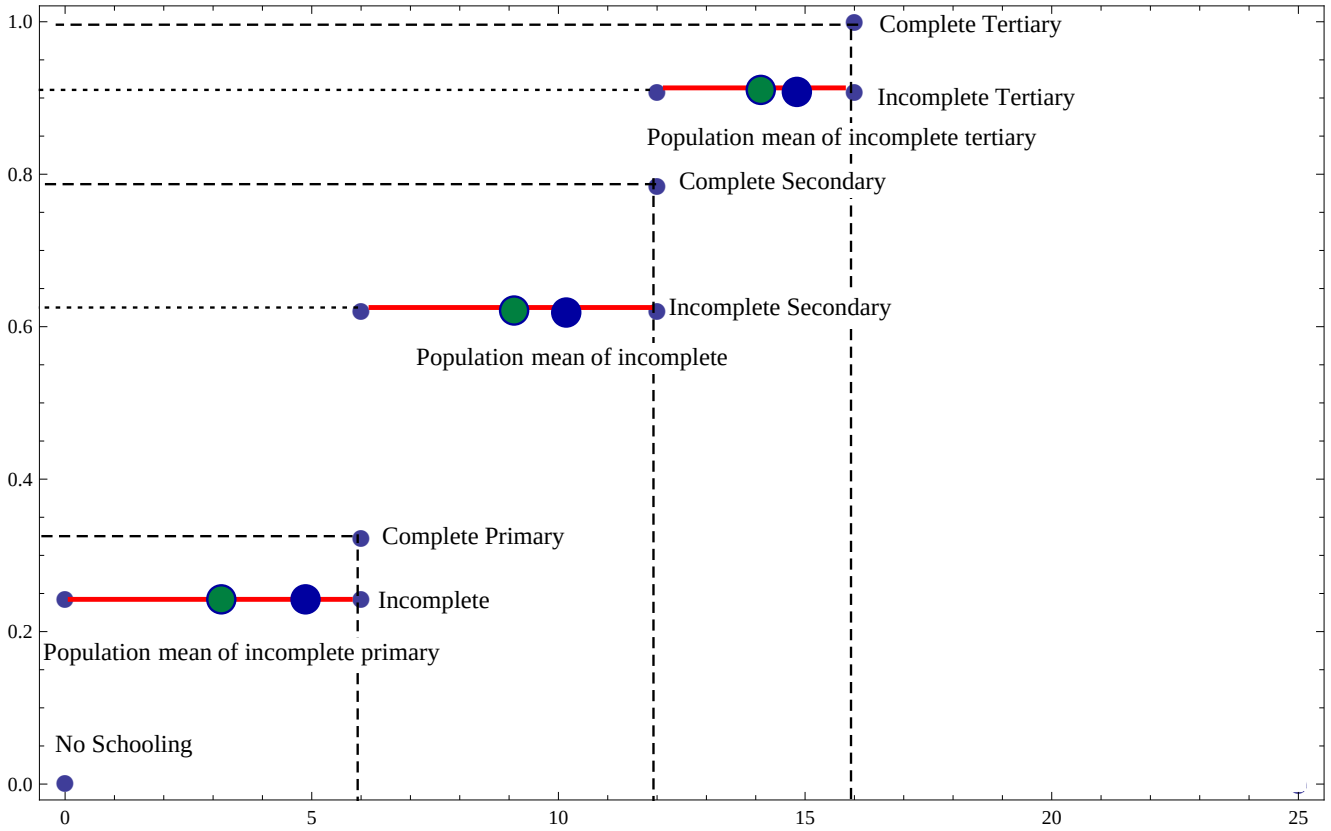


Example: Japan (1990)



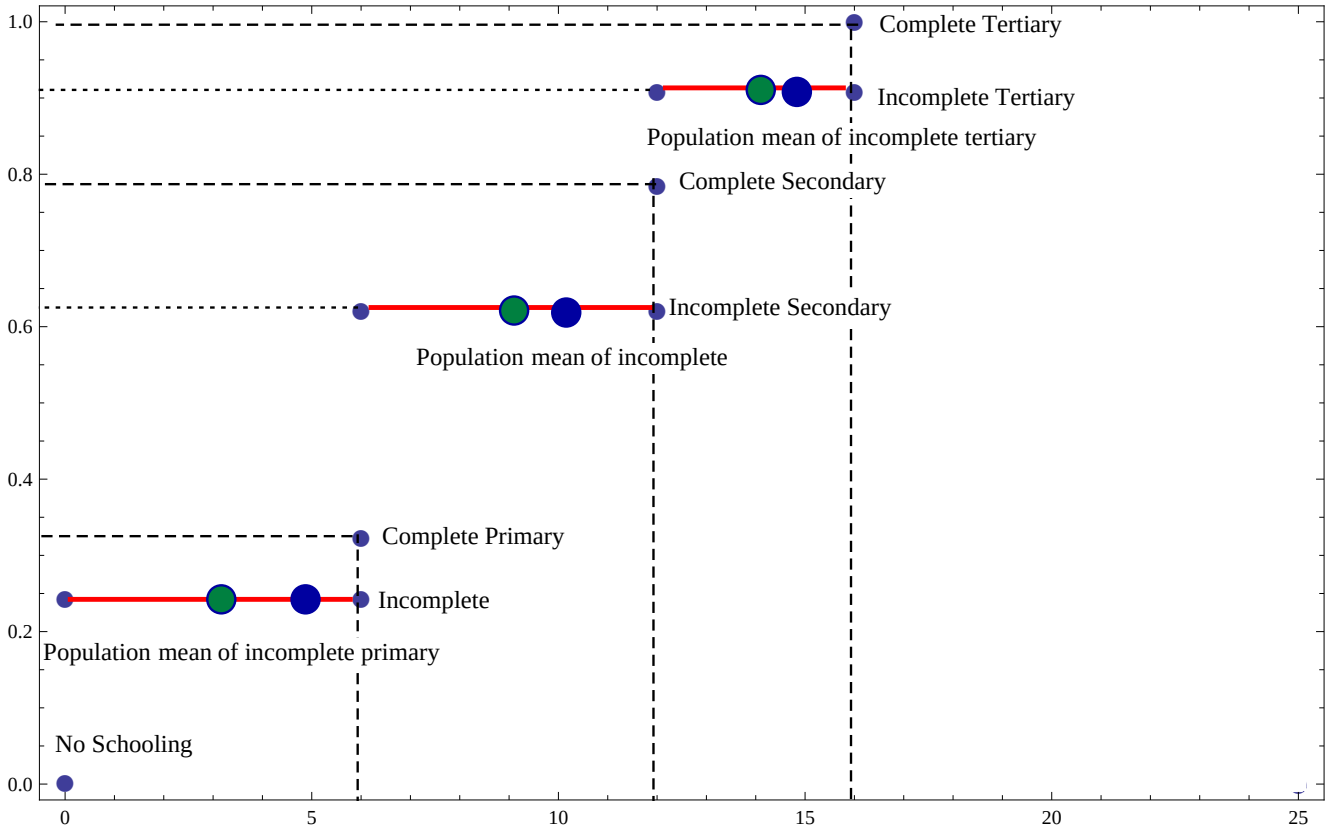
It is not possible to approximate the direction of the bias

Example: Japan (1990)



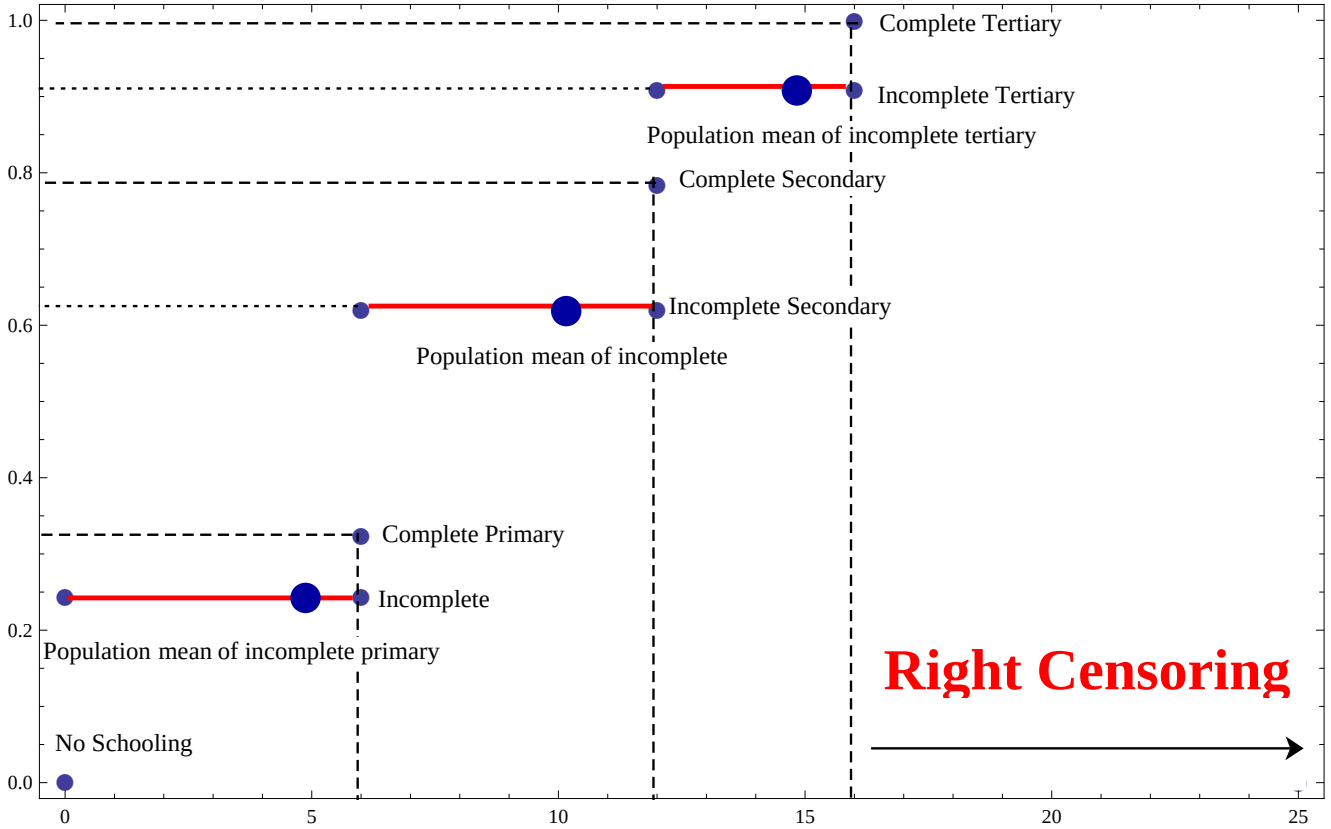
In this case UNESCO's methodology would underestimate the MYS

Example: Japan (1990)



In this case UNESCO's methodology would underestimate the MYS

Example: Japan (1990)



Even if we had information on the mean of incomplete levels, MYS would be biased

Inequality in education

- The distribution of educational outcomes has also been of main interest in the literature.
- International inequality in education: a lower bound? (e.g. Ram,1990; World Bank, 2005)
- National inequality using the same approach as for the MYS. (Castello and Domenech, 2008, Thomas et al., 2001, Meschi and Scervini, 2013)
- The same sources of bias
  - No information on incomplete levels
  - Right censoring
  - **ADITIONALLY:** differences within each educational level are not considered.

Inequality in education

- The distribution of educational outcomes has also been of main interest in the literature.
- International inequality in education: a lower bound? (e.g. Ram,1990; World Bank, 2005)
- National inequality using the same approach as for the MYS. (Castello and Domenech, 2008, Thomas et al., 2001, Meschi and Scervini, 2013)
- The same sources of bias
  - No information on incomplete levels
  - Right censoring
  - **ADITIONALLY:** differences within each educational level are not considered.

Inequality in education

- The distribution of educational outcomes has also been of main interest in the literature.
- International inequality in education: a lower bound? (e.g. Ram,1990; World Bank, 2005)
- National inequality using the same approach as for the MYS. (Castello and Domenech, 2008, Thomas et al., 2001, Meschi and Scervini, 2013)
- The same sources of bias
  - No information on incomplete levels
  - Right censoring
  - **ADITIONALLY:** differences within each educational level are not considered.

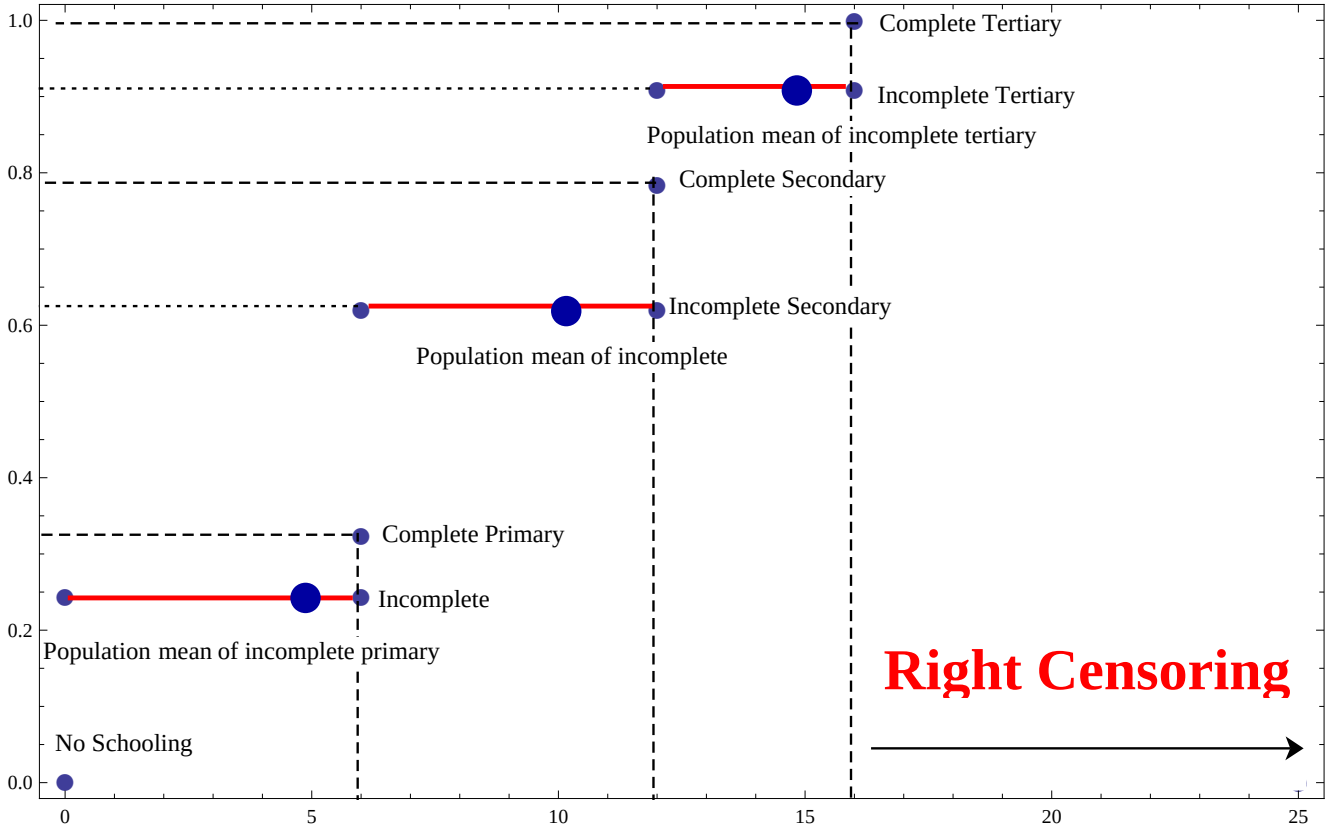
Inequality in education

- The distribution of educational outcomes has also been of main interest in the literature.
- International inequality in education: a lower bound? (e.g. Ram,1990; World Bank, 2005)
- National inequality using the same approach as for the MYS. (Castello and Domenech, 2008, Thomas et al., 2001, Meschi and Scervini, 2013)
- The same sources of bias
  - No information on incomplete levels
  - Right censoring
  - **ADITIONALLY:** differences within each educational level are not considered.

Inequality in education

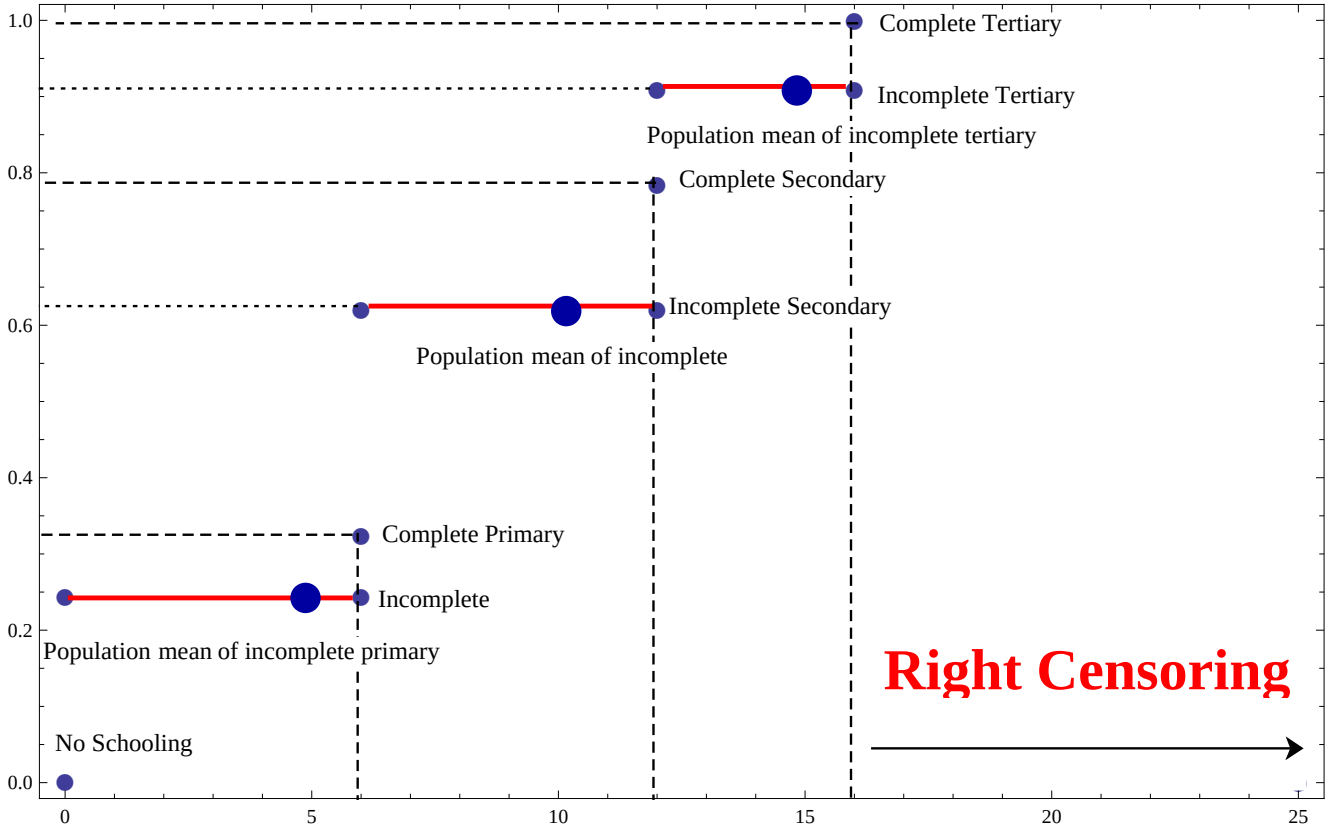
- The distribution of educational outcomes has also been of main interest in the literature.
- International inequality in education: a lower bound? (e.g. Ram,1990; World Bank, 2005)
- National inequality using the same approach as for the MYS. (Castello and Domenech, 2008, Thomas et al., 2001, Meschi and Scervini, 2013)
- The same sources of bias
  - No information on incomplete levels
  - Right censoring
  - **ADITIONALLY:** differences within each educational level are not considered.

Example: Japan (1990)



Even if we had information on the mean of incomplete levels, MYS would be biased

Example: Japan (1990)



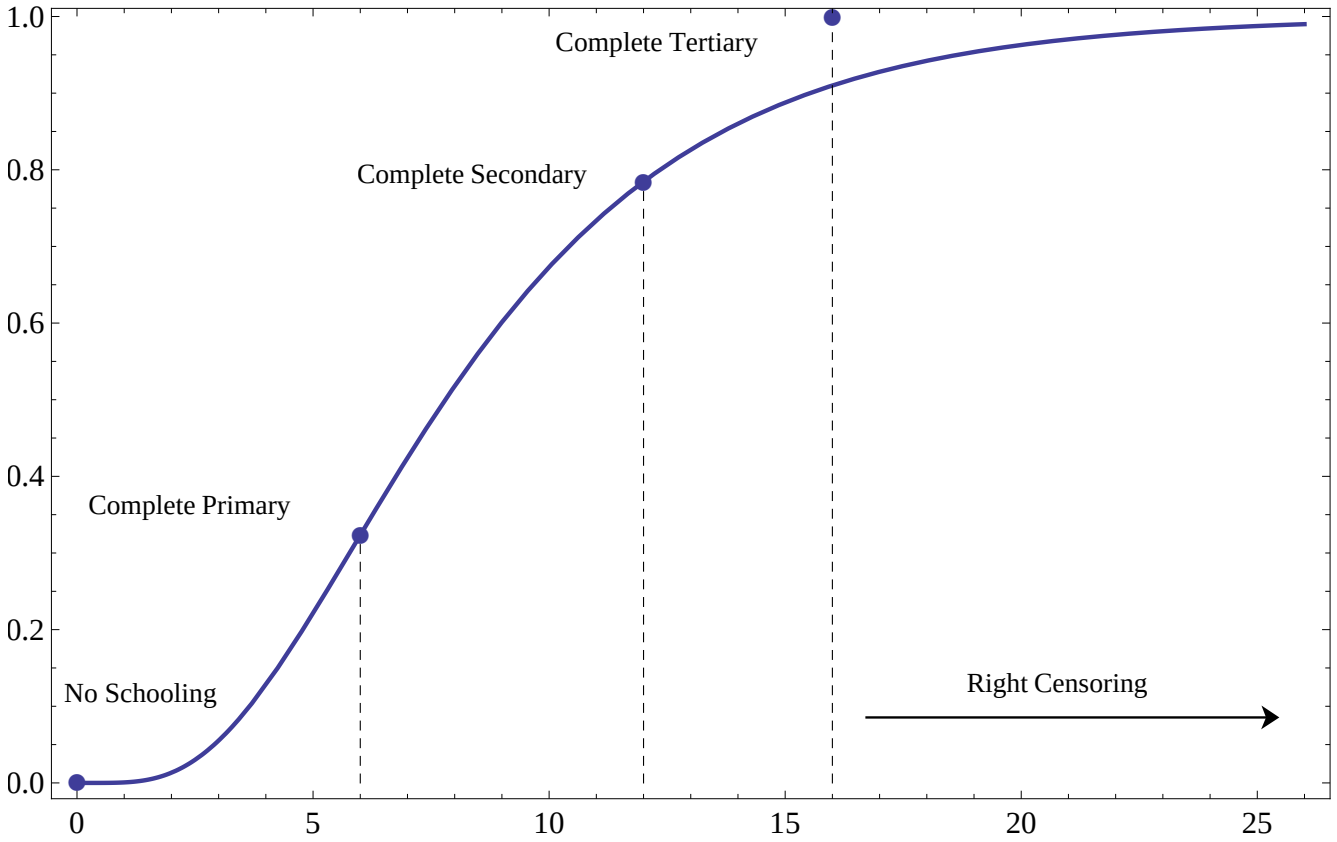
Even if we had information on the mean of incomplete levels, MYS would be biased

Methodology

MAIN CHALLENGE: Combine grouped information on educational attainment and duration levels to obtain reliable estimates of the distribution of schooling.

- Educational attainment rates of each educational level represent points of the CDF.
- Parametric model to link the empirical points of CDF.

Example: Japan (1990)



Methodology

MAIN CHALLENGE: Combine grouped information on educational attainment and duration levels to obtain reliable estimates of the distribution of schooling.

- Main limitation: the functional form of the distribution is defined ex-ante.
- Misspecification of the model leads inconsistent estimates.
- It is not possible to test the adequacy of this model for the unknown parts of the distribution (within educational stages).
- If we are able to fit the available information with high precision, there is no reason to assume that our model leads poorer estimates than those can be found in the literature.

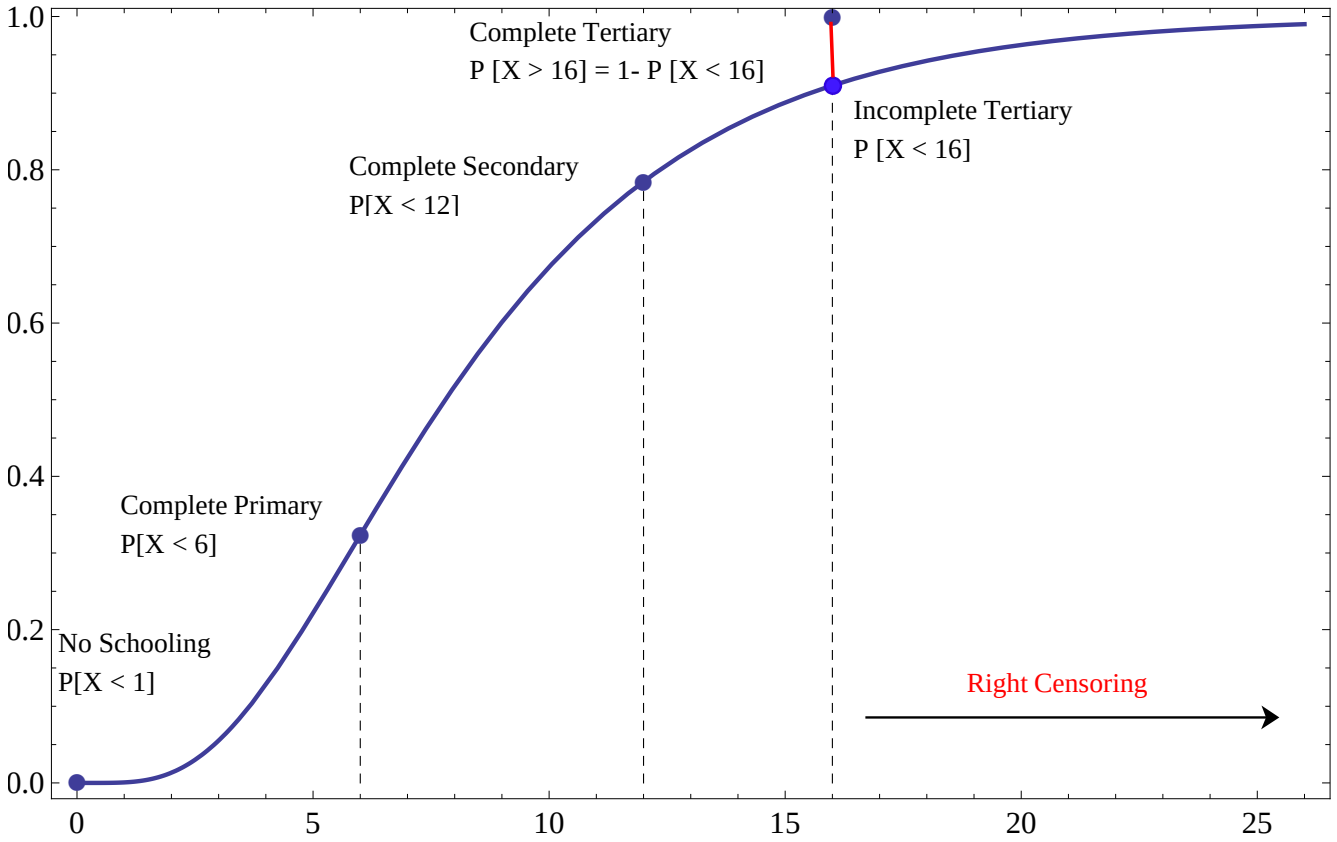
Data

- BL data on educational attainments of the population over 15 ($h_i^{(j)}$).
  - No schooling
  - Primary
  - Secondary
  - Tertiary (Right censoring)
    - complete
    - incomplete
- Data on the official duration of primary and secondary education is drawn from UNESCO ($D_i^{(j)}$).

Data

- BL data on educational attainments of the population over 15 ($h_i^{(j)}$).
  - No schooling $<$ 1 year
  - Primary official duration
  - Secondary official duration
  - Tertiary (Right censoring)
    - complete $>$ 4 years
    - incomplete$<$4 years
- Data on the official duration of primary and secondary education is drawn from UNESCO ($D_i^{(j)}$).

Example: Japan (1990)



How to chose a model for the distribution of education?

Definition of the educational variable

Let X be a continuous variable representing the time of schooling until either completing the maximum level of education or dropping out the school.

- Survival analysis provides an optimal framework.
- The same parametric functional form is used for all countries over the whole period.
- We need a flexible model due to the high degree of heterogeneity.
  - Global analysis of 143 countries.
  - Long period of time: 1970-2010.

How to chose a model for the distribution of education?

Definition of the educational variable

Let X be a continuous variable representing the time of schooling until either completing the maximum level of education or dropping out the school.

- Survival analysis provides an optimal framework.
- The same parametric functional form is used for all countries over the whole period.
- We need a flexible model due to the high degree of heterogeneity.
  - Global analysis of 143 countries.
  - Long period of time: 1970-2010.

How to chose a model for the distribution of education?

Definition of the educational variable

Let X be a continuous variable representing the time of schooling until either completing the maximum level of education or dropping out the school.

GENERALIZED GAMMA DISTRIBUTION

- Includes most of the distributions commonly used in survival analysis. (the Weibull, the exponential and the gamma distributions)
- It is able to represent one and zero mode distributions.
  - One mode is expected in developed countries with compulsory years of schooling
  - Zero mode distributions are characteristic of developing countries, which present high illiteracy rates.

How to chose a model for the distribution of education?

Definition of the educational variable

Let X be a continuous variable representing the time of schooling until either completing the maximum level of education or dropping out the school.

GENERALIZED GAMMA DISTRIBUTION

- Includes most of the distributions commonly used in survival analysis. (the Weibull, the exponential and the gamma distributions)
- It is able to represent one and zero mode distributions.
  - One mode is expected in developed countries with compulsory years of schooling
  - Zero mode distributions are characteristic of developing countries, which present high illiteracy rates.

Estimation Methods

Nonlinear least squares

$$SSR = \sum_{j=1}^{J-1} \left(IG(u_{it}^{(j)}; p, \beta^a) - h_{it}^{(j)} \right)^2 + \left(1 - IG(u_{it}^{(TC)}; p, \beta^a) - h_{it}^{(TC)} \right)^2 + \epsilon_{it},$$

- $(IG(u_{it}^{(j)}; p, \beta^a)$ CDF of the generalized gamma distribution
  - $u_{it} = \left(D_{it}^{(j)} \right)^a$.
  - IG(.) stands for the incomplete gamma function.
- $h_{it}^{(j)}$ Observed attainment rates.
- ϵ_i is the error term.
- The second term accounts for right censoring.

Estimation Methods

Nonlinear least squares

$$SSR = \sum_{j=1}^{J-1} \left(IG(u_{it}^{(j)}; p, \beta^a) - h_{it}^{(j)} \right)^2 + \left(1 - IG(u_{it}^{(TC)}; p, \beta^a) - h_{it}^{(TC)} \right)^2 + \epsilon_{it},$$

- $(IG(u_{it}^{(j)}; p, \beta^a)$ CDF of the generalized gamma distribution
  - $u_{it} = \left(D_{it}^{(j)} \right)^a$.
  - IG(.) stands for the incomplete gamma function.
- $h_{it}^{(j)}$ Observed attainment rates.
- ϵ_i is the error term.
- The second term accounts for right censoring.

Estimation Methods

Nonlinear least squares

$$SSR = \sum_{j=1}^{J-1} \left(IG(u_{it}^{(j)}; p, \beta^a) - h_{it}^{(j)} \right)^2 + \left(1 - IG(u_{it}^{(TC)}; p, \beta^a) - h_{it}^{(TC)} \right)^2 + \epsilon_{it},$$

- $(IG(u_{it}^{(j)}; p, \beta^a)$ CDF of the generalized gamma distribution
  - $u_{it} = \left(D_{it}^{(j)} \right)^a$.
  - IG(.) stands for the incomplete gamma function.
- $h_{it}^{(j)}$ Observed attainment rates.
- ϵ_i is the error term.
- The second term accounts for right censoring.

Construction of the database

- We fit the GG for each country (i = 1, ..., 143) from 1970 to 2010 at five year intervals.
- MYS is given by the mean of the distribution,

$$\mu = \frac{\beta\Gamma(p + \frac{1}{a})}{\Gamma(p)}.$$

- GE measures are given by the following expressions (Jenkins, 2009; Sarabia et al., 2015)

$$GE(2) = -\frac{1}{2} + \frac{\Gamma(p + \frac{2}{a})\Gamma(p)}{2\Gamma^2(p + \frac{1}{a})},$$

$$T_0 = -\frac{\psi(p)}{a} + \log\frac{\Gamma(p + \frac{1}{a})}{\Gamma(p)},$$

$$T_1 = \frac{\psi(p + \frac{1}{a})}{a} - \log\frac{\Gamma(p + \frac{1}{a})}{\Gamma(p)}.$$

Construction of the database

- We fit the GG for each country ($i = 1, ..., 143$) from 1970 to 2010 at five year intervals.
- MYS is given by the mean of the distribution,

$$\mu = \frac{\beta\Gamma(p + \frac{1}{a})}{\Gamma(p)}.$$

- GE measures are given by the following expressions (Jenkins, 2009; Sarabia et al., 2015)

$$GE(2) = -\frac{1}{2} + \frac{\Gamma(p + \frac{2}{a})\Gamma(p)}{2\Gamma^2(p + \frac{1}{a})},$$

$$T_0 = -\frac{\psi(p)}{a} + \log\frac{\Gamma(p + \frac{1}{a})}{\Gamma(p)},$$

$$T_1 = \frac{\psi(p + \frac{1}{a})}{a} - \log\frac{\Gamma(p + \frac{1}{a})}{\Gamma(p)}.$$

Construction of the database

- We fit the GG for each country ($i = 1, ..., 143$) from 1970 to 2010 at five year intervals.
- MYS is given by the mean of the distribution,

$$\mu = \frac{\beta\Gamma(p + \frac{1}{a})}{\Gamma(p)}.$$

- GE measures are given by the following expressions (Jenkins, 2009; Sarabia et al., 2015)

$$GE(2) = -\frac{1}{2} + \frac{\Gamma(p + \frac{2}{a})\Gamma(p)}{2\Gamma^2(p + \frac{1}{a})},$$

$$T_0 = -\frac{\psi(p)}{a} + \log\frac{\Gamma(p + \frac{1}{a})}{\Gamma(p)},$$

$$T_1 = \frac{\psi(p + \frac{1}{a})}{a} - \log\frac{\Gamma(p + \frac{1}{a})}{\Gamma(p)}.$$

The global distribution of education

Mixture of national distributions

Let $X_i, i = 1, ..., N$ be the educational attainment in the county i which is assumed to follow a GG distribution. The global CDF can be expressed as,

$$F(x) = \sum_{i=1}^{N} \lambda_i F_i(x) = \sum_{i=1}^{N} \lambda_i IG(u; p_i, \beta_i^{a_i}),$$

where λ_i stands for the population weights of the countries.

Global MYS

The mean of the mixture is ivan by,

$$\mu = \sum_{i=1}^{N} \lambda_i \mu_i = \sum_{i=1}^{N} \lambda_i \frac{\beta_i \Gamma(p_i + \frac{1}{a_i})}{\Gamma(p_i)}.$$

The global distribution of education

Mixture of national distributions

Let $X_i, i = 1, ..., N$ be the educational attainment in the county i which is assumed to follow a GG distribution. The global CDF can be expressed as,

$$F(x) = \sum_{i=1}^{N} \lambda_i F_i(x) = \sum_{i=1}^{N} \lambda_i IG(u; p_i, \beta_i^{a_i}),$$

where λ_i stands for the population weights of the countries.

Global MYS

The mean of the mixture is ivan by,

$$\mu = \sum_{i=1}^{N} \lambda_i \mu_i = \sum_{i=1}^{N} \lambda_i \frac{\beta_i \Gamma(p_i + \frac{1}{a_i})}{\Gamma(p_i)}.$$

Global inequality in education

Generalized Entropy measures

$$I(\theta) = \sum_{i=1}^{N} \lambda_i^{1-\theta} s_i^{\theta} I_i^{(\theta)} + \frac{1}{\theta(\theta-1)} \left(\sum_{i=1}^{N} \lambda_i^{1-\theta} \left(\frac{\mu_i}{\mu} \right)^{\theta} - 1 \right).$$

λ_i and $I_i^{(\theta)}$ are the population share and the GE measure of the country i. s_i stands for the proportion of MYS of the country i in the global mean.

Theil's Entropy index

$$T_W = \sum_{i=1}^{N} s_i T_i; T_B = \sum_{i=1}^{N} s_i \log \left(\frac{\mu_i}{\mu} \right),$$

Mean Log Deviation

$$L_W = \sum_{i=1}^{N} \lambda_i L_i; L_B = \sum_{i=1}^{N} \lambda_i \log \left(\frac{\mu}{\mu_i} \right),$$

Global inequality in education

Generalized Entropy measures

$$I(\theta) = \sum_{i=1}^{N} \lambda_i^{1-\theta} s_i^{\theta} I_i^{(\theta)} + \frac{1}{\theta(\theta-1)} \left(\sum_{i=1}^{N} \lambda_i^{1-\theta} \left(\frac{\mu_i}{\mu} \right)^{\theta} - 1 \right).$$

λ_i and $I_i^{(\theta)}$ are the population share and the GE measure of the country i. s_i stands for the proportion of MYS of the country i in the global mean.

Theil's Entropy index

$$T_W = \sum_{i=1}^{N} s_i T_i; T_B = \sum_{i=1}^{N} s_i \log \left(\frac{\mu_i}{\mu} \right),$$

Mean Log Deviation

$$L_W = \sum_{i=1}^{N} \lambda_i L_i; L_B = \sum_{i=1}^{N} \lambda_i \log \left(\frac{\mu}{\mu_i} \right),$$

Global inequality in education

Generalized Entropy measures

$$I(\theta) = \sum_{i=1}^{N} \lambda_i^{1-\theta} s_i^{\theta} I_i^{(\theta)} + \frac{1}{\theta(\theta-1)} \left(\sum_{i=1}^{N} \lambda_i^{1-\theta} \left(\frac{\mu_i}{\mu} \right)^{\theta} - 1 \right).$$

λ_i and $I_i^{(\theta)}$ are the population share and the GE measure of the country i. s_i stands for the proportion of MYS of the country i in the global mean.

Theil's Entropy index

$$T_W = \sum_{i=1}^{N} s_i T_i; T_B = \sum_{i=1}^{N} s_i \log \left(\frac{\mu_i}{\mu} \right),$$

Mean Log Deviation

$$L_W = \sum_{i=1}^{N} \lambda_i L_i; L_B = \sum_{i=1}^{N} \lambda_i \log \left(\frac{\mu}{\mu_i} \right),$$

Goodness of fit

Table: Empirical and estimated educational attainment rates.

Year		NS	PS or less	SS or less	TS or more	SSD
1970	GD	1.0000	5.1706	11.1916	15.1916	
	BL	0.3495	0.7357	0.9612	0.0388	
	Estimated %	0.3476	0.7470	0.9560	0.0440	0.0002
1980	GD	1.0000	5.1591	11.6410	15.6410	
	BL	0.2949	0.6307	0.9383	0.0618	
	Estimated %	0.2900	0.6498	0.9240	0.0760	0.0008
1990	GD	1.0000	5.1775	11.6283	15.6283	
	BL	0.2559	0.5613	0.9095	0.0905	
	Estimated %	0.2429	0.5822	0.8969	0.1031	0.0009
2000	GD	1.0000	5.2219	11.6524	15.6524	
	BL %	0.1947	0.4603	0.8784	0.1217	
	Estimated %	0.1730	0.4857	0.8635	0.1365	0.0016
2010	GD	1.0000	5.2814	11.7268	15.7268	
	BL	0.1435	0.3658	0.8587	0.1413	
	Estimated %	0.1159	0.3856	0.8341	0.1659	0.0024

Comparison with Barro and Lee (2013) data

Table: Comparison of BL data on average years of schooling (AYS-BL) with the theoretical values of MYS for the GG distribution.

Year	Levels		Change		Correlation		
	AYS-BL	MYS	AYS-BL	MYS	levels	5-year D.	10-year D.
1970	4.4564	3.4212	-	-	0.9097	-	-
1975	4.8834	3.8937	9.58%	13.81%	0.9142	0.8215	-
1980	5.3777	4.4311	10.12%	13.80%	0.9260	0.7668	0.8534
1985	5.7794	4.8776	7.47%	10.07%	0.9301	0.7794	0.7603
1990	6.1446	5.2060	6.32%	6.73%	0.9356	0.8050	0.8090
1995	6.6993	5.7417	9.03%	10.29%	0.9487	0.8341	0.8262
2000	7.2248	6.2320	7.84%	8.54%	0.9536	0.6351	0.7893
2005	7.6943	6.7063	6.50%	7.61%	0.9587	0.7747	0.7463
2010	8.0920	7.0870	5.17%	5.68%	0.9612	0.8662	0.7536

Comparison with Barro and Lee (2013) data

Table: Comparison of BL data on average years of schooling (AYS-BL) with the theoretical values of MYS for the GG distribution.

Year	Levels		Change		Correlation		
	AYS-BL	MYS	AYS-BL	MYS	levels	5-year D.	10-year D.
1970	4.4564	3.4212	-	-	0.9097	-	-
1975	4.8834	3.8937	9.58%	13.81%	0.9142	0.8215	-
1980	5.3777	4.4311	10.12%	13.80%	0.9260	0.7668	0.8534
1985	5.7794	4.8776	7.47%	10.07%	0.9301	0.7794	0.7603
1990	6.1446	5.2060	6.32%	6.73%	0.9356	0.8050	0.8090
1995	6.6993	5.7417	9.03%	10.29%	0.9487	0.8341	0.8262
2000	7.2248	6.2320	7.84%	8.54%	0.9536	0.6351	0.7893
2005	7.6943	6.7063	6.50%	7.61%	0.9587	0.7747	0.7463
2010	8.0920	7.0870	5.17%	5.68%	0.9612	0.8662	0.7536

The global distribution of education

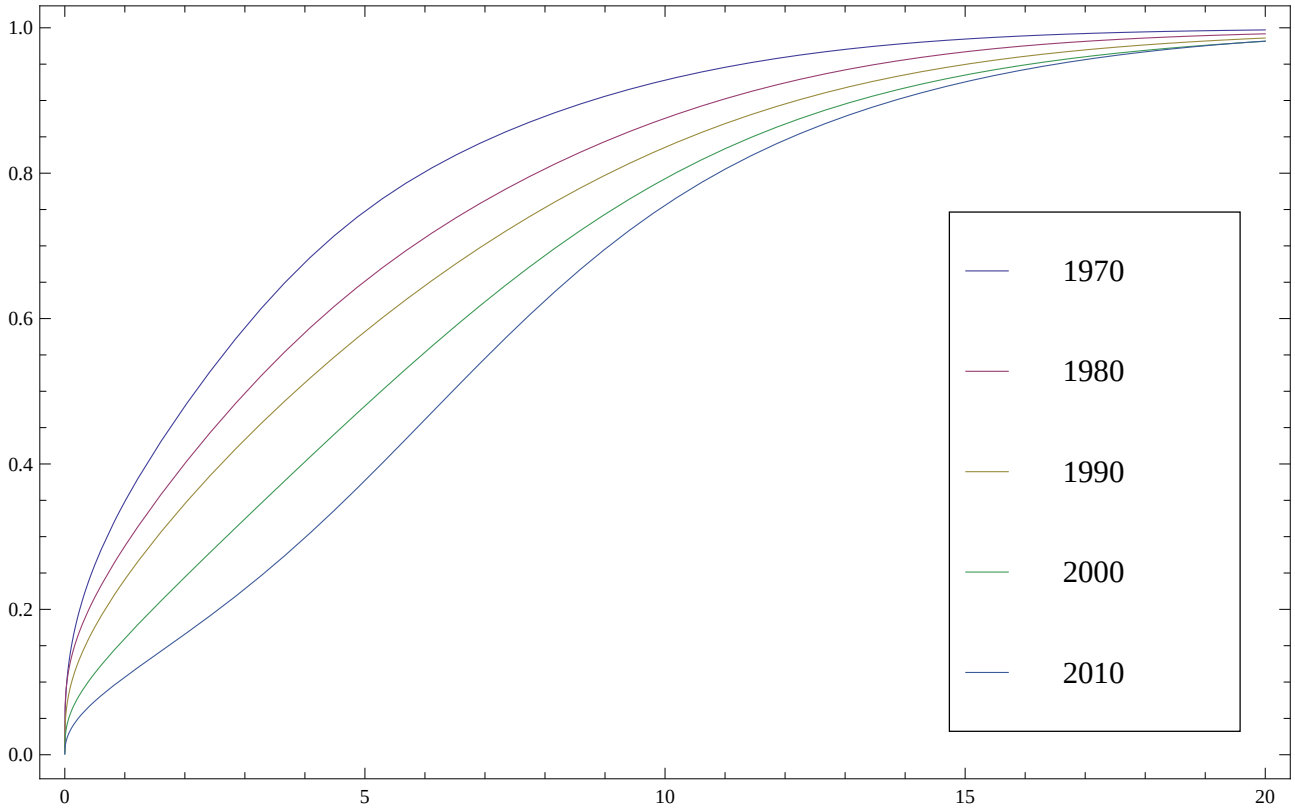


Figure: Global CDF of years of schooling

The global distribution of education

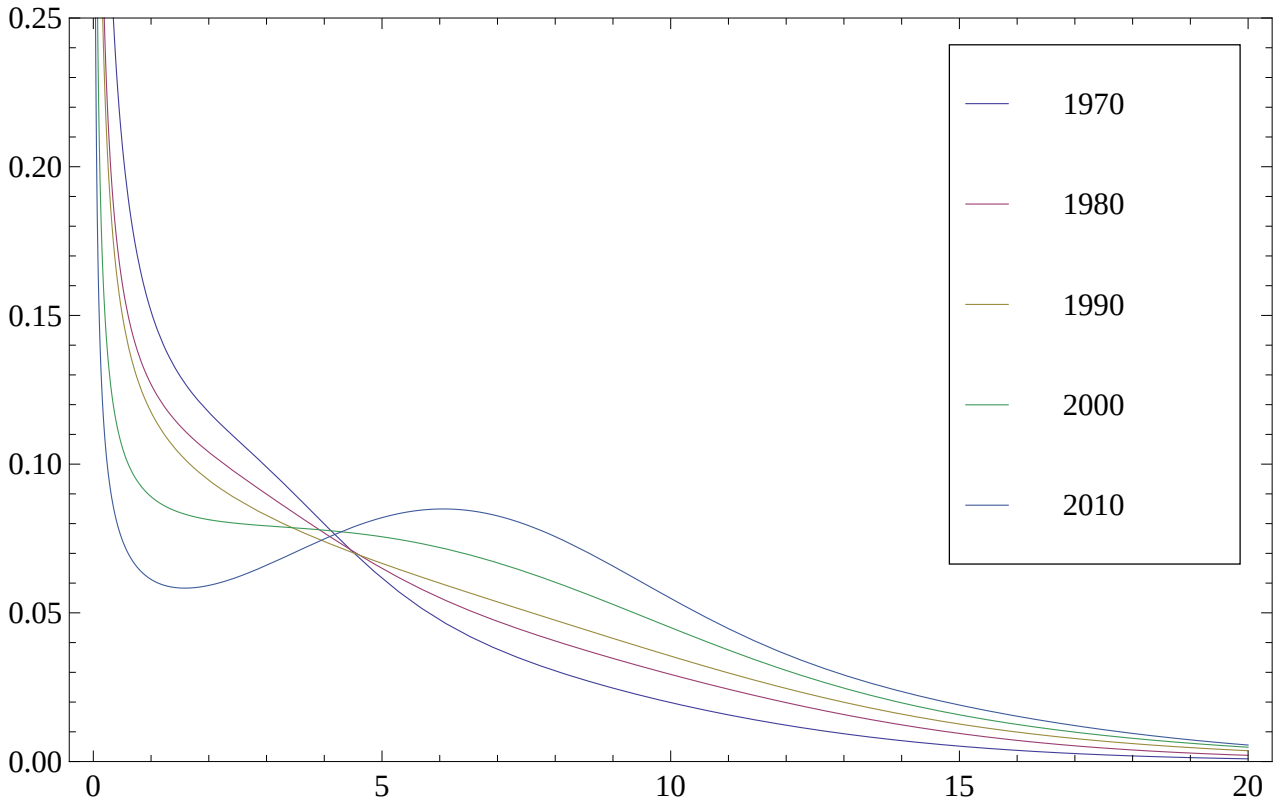


Figure: Global PDF of years of schooling

Global inequality in education

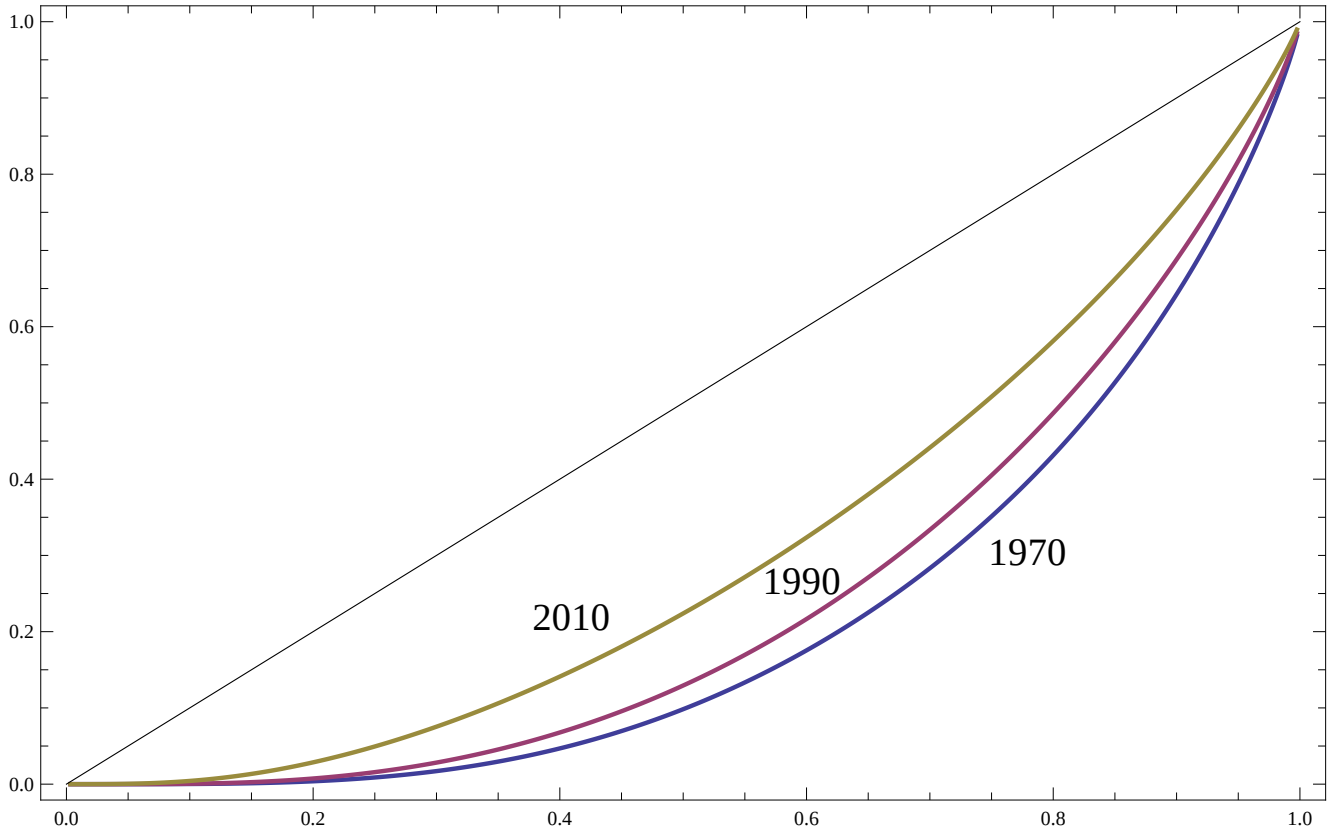


Figure: Global Lorenz curves of education

Global inequality in education

Table: Global inequality in education (1970-2010).

Year	1970	1975	1980	1985	1990	1995	2000	2005	2010
MLD	1.4711	1.5942	1.4829	1.2129	1.0990	0.9134	0.7611	0.6110	0.5232
Between	0.1945	0.1645	0.1339	0.1131	0.1054	0.0920	0.0790	0.0715	0.0610
Within	1.2766	1.4297	1.3490	1.0998	0.9936	0.8214	0.6821	0.5395	0.4622
% Between	13.22%	10.32%	9.03%	9.33%	9.59%	10.07%	10.37%	11.71%	11.67%
% Within	86.78%	89.68%	90.97%	90.67%	90.41%	89.93%	89.63%	88.29%	88.33%

Conclusions

- Existing databases include biased the estimates of MYS and inequality measures.
- We have used a flexible parametric model to estimate the distribution of years of schooling.
- In global terms, average educational outcomes by BL are substantially higher than our estimates.
- We have also focused on the global distribution of education.
  - Expansion of all levels of education.
  - Reduction of inequality levels.
  - Differences across countries represented a reduced proportion of global inequality.

Conclusions

- Existing databases include biased the estimates of MYS and inequality measures.
- We have used a flexible parametric model to estimate the distribution of years of schooling.
- In global terms, average educational outcomes by BL are substantially higher than our estimates.
- We have also focused on the global distribution of education.
  - Expansion of all levels of education.
  - Reduction of inequality levels.
  - Differences across countries represented a reduced proportion of global inequality.

Conclusions

- Existing databases include biased the estimates of MYS and inequality measures.
- We have used a flexible parametric model to estimate the distribution of years of schooling.
- In global terms, average educational outcomes by BL are substantially higher than our estimates.
- We have also focused on the global distribution of education.
  - Expansion of all levels of education.
  - Reduction of inequality levels.
  - Differences across countries represented a reduced proportion of global inequality.

Conclusions

- Existing databases include biased the estimates of MYS and inequality measures.
- We have used a flexible parametric model to estimate the distribution of years of schooling.
- In global terms, average educational outcomes by BL are substantially higher than our estimates.
- We have also focused on the global distribution of education.
  - Expansion of all levels of education.
  - Reduction of inequality levels.
  - Differences across countries represented a reduced proportion of global inequality.