Can The Distribution of Highest Educational Attainment Be Characterised by A Discrete Probability Distribution?

Debasis Bandyopadhyay
Department of Economics
The University of Auckland
Private Bag 92019, Auckland, New Zealand
E-mail: debasis@auckland.ac.nz

Abstract

Distribution of human capital is a recent addition in the literature to the list of a few fundamental determinants of growth. This paper addresses an important problem associated with the empirical characterization of that distribution by utilizing the recently available distribution of the highest educational attainment in the labor force. We tried to fit the distribution of the number of years of school by over-dispersed Poisson and Negative-Binomial distributions. Based on the data compiled by Barro and Lee, none of the discrete distributions fit the data. The standard discrete probability distributions are too smooth to account for the important information contained in the data, ie., schooling is more likely to be terminated at the completion of a category of schooling (e.g., primary, secondary, and higher education) than during a category, an important feature contained in the frequency distribution of highest educational attainment. Future research of this data should focus on more complex models which account for this "discontinuity" in the data, or modeling of other important features of the frequency distribution of highest educational attainment.

1. Introduction

Economic growth is one of the essential concerns in macroeconomics. Economists are still trying to understand why countries experience sharp divergence in long-term per-capita income and growth rates with a resulting experience of dramatically different standards of living.

The neoclassical theory sees the accumulation of capital as the primary force that leads an economy to reach its steady-state, but attributes technological progress as the driving force behind long-term growth. Within the scope of endogenous growth literature, one strand of theory stresses the role that capital accumulation - with a broader interpretation of capital that includes human capital - plays in determining long-term growth. The second approach casts external economies in a leading role in the growth process; and the third branch of the theory incorporates intentional investments of resources by profit-seeking firms in order to generate technological improvements. These different lines of thought share a common history: the growth theory of Solow (1956). What matters is the power of the theory to explain the stylized facts in recent growth experiences. The impact of human capital on growth was excluded in the standard Solow model, but was treated as an influential factor in subsequent studies. In the past decade, the contribution of human capital in either the neoclassical framework or in an endogenous growth model have been analyzed extensively. Mankiw, Romer, and Weil (1992) made a significant contribution by augmenting the standard Solow model to incorporate human capital. A broader definition of capital strengthens the predictions in the Solow model, and thereby advocates its validity in the field. On the other hand, the flexibility of the endogenous growth theory allows the models to stress the role of human

capital in various ways. Romer (1986, 1989, 1990) and Grossman and Helpman (1991) treat human capital as a critical input in the production of new knowledge or designs while Lucas (1988) assumes that average human capital is a proxy for the spillover effect of technology progress. In these models, human capital is the driving force of growth. Recently, Galor (1994) and others have modeled the distribution of human capital as a fundamental variable of economic growth.

Bandyopadhyay (1993) combines discrete occupational choice of Banerjee and Newman and the endogenous growth model of Lucas (1988) to generate a theoretical framework where the initial distribution of human capital is a crucial explanatory factor of a country's long term growth rate and income distribution. The model shows numerically for artificially simulated economies that different initial distributions of human capital could lead to different dynamics of economic growth and income inequality. In particular, according to the model world described in the thesis, the striking contrast between the high growth of East Asia and the slow growth of the Latin America could be rationalized by a special conjecture. The conjecture is that the initial distribution of human capital in East Asia was more equitable than Latin America. For other conjecture involving a concept of human capital distribution see also Chari and Hopenhayn (1991), Galor and Tsiddon (1994) and Togo (1996). Those literature on income distribution and growth based on human capital theory add a new dimension to macroeconomic dynamics by making the distribution of human capital a fundamental determinant of the macroeconomic aggregates.

The theoretical models mentioned earlier, and the literature in general, demand but do not provide systematic or stylized observations on an international comparison of

distribution of human capital over time. Robert M. Solow, a Nobel Laureate in economics, encouraged future researchers to fill that vacuum in the literature in the 1992 George Seltzer Distinguished Lecture Series at the University of Minnesota entitled *Growth with Equity with Investment in Human Capital.* Several attempts have been made to compile data on human capital distribution. Mincer (1991) and Krueger (1993) are examples of work related to gathering data on human capital distribution to be used as evidence on theoretical models. They are, however, mainly concerned with the US data.

Barro and Lee (1993) compiled data on highest educational attainment, one of the measures of human capital, among adult population (25 and older) for a broad cross section of countries. The data was given over five-year intervals from 1960 to 1985. The data provides the fractions of population belonging to seven categories: no formal education (NQ), incomplete primary, complete primary, first cycle of secondary, second cycle of secondary, incomplete higher, and complete higher (HQ). They created the data using census information on school attainment for adult population which were obtained from UNESCO publication and other sources. School enrollment ratios were used to fill in the missing observations. See Barro and Lee (1997) for an update of their data set for the population aged 15 and over. Nehru, Swanson and Dubey (1995) created a series of estimates of stock of education in 85 countries over 28 years (1960-87) for the population between the ages 15 and 64. They used enrollment data from UNESCO sources and corrected their estimates for grade repetition among school-goers and country specific drop-out rates for primary and secondary students. Prior to Barro and Lee (1993, 1997) and Nehru, Swanson and Dubey (1995) there were several studies on the international comparisons of various measures of human capital. A few notable papers in this area of

research include among others Psacharopoulos and Arriagada (1986), Lau, et. al (1993) and Kaneko (1986).

The average years of schooling has been increasing in almost all countries since the 60s (see, e.g., Barro and Lee, 1997), but only a small group of countries has been enjoying more than 3% annual average growth rate between 1965-90. The following quotation (see the web page maintained by the National Bureau of Economic Research, Inc., http://www.nber.org/programs/efg/efg.html) is in conformity with the preliminary research conducted by the Bandyopadhyay (1997). "The first meeting of the newly formed "Growth Group" focused on the accumulation and development of human capital, finding some surprisingly paradoxical results and developing exciting avenues for future research. Lant Pritchett of the World Bank presented cross-sectional evidence that the growth of human capital, as measured by years of education, is completely uncorrelated with the growth of output. This result is surprisingly robust to the use of different data sets, as confirmed by conference participant Jong-Wha Lee, NBER and Korea University who, together with Robert J. Barro, NBER and Harvard University, has developed a broad international database on education. The conventional measure of human capital, the years that students devote to education, is extraordinarily crude, providing inadequate assessment of the value and growth of human capital...."

Most of the current research indicates that the average years of schooling is not a good measure of human capital. In order to come up with a better measure of human capital, it is important to explore the following question: Is it possible to describe the highest educational attainment distributions for different countries by a well-known

probability distribution with possibly different parameters? In this paper, we investigate this important problem using an existing database created by Barro and Lee (1993).

2. Research Hypothesis

Existing data (Barro & Lee 1993) gives the proportion of the population over 25 years of age in each country in various educational categories based on their highest education attainment. Additional data is available from other sources on the population over 25 years in each country, making it possible to reconstruct the observed frequencies in each educational category.

One possible model for this data is to assume that each individuals total number of years of schooling is a drawing from a random distribution (F_1), with different parameters for each country. These parameters, in turn, are drawn from a random distribution (F_2) with fixed parameters for a set group of countries (*i.e.*, top income quintile, bottom income quintile, etc). A version of this model that could be appropriate to the human capital distribution data is with F_1 the Poisson distribution, and F_2 the Gamma distribution.

Let X_i be a random variable representing the highest educational attainment among the 25+ age group for the *i*th country. We shall consider the following two multi-level models:

Model 1 (Poisson-Gamma Model):

Let F_1 be the Poisson distribution, and F_2 be the Beta distribution.

(i) $X_i \mid \lambda_I \sim$ Poisson (λ_i), $\quad i=1,\ldots,m;$

(ii) *A priori* $\lambda_i \sim$ Gamma (α,β), $\quad i=1,\ldots,m;$

where the Gamma distribution has the following pdf

$$f(\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} e^{-\beta\lambda} \lambda^{\alpha-1}, \lambda > 0 .$$

Model 2 (Negative Binomial- Beta Model):

Let F_1 be the Negative Binomial distribution, and F_2 be the Beta distribution.

(i) $X_i \mid p_i \sim$ Negative Binomial (α_i, p_i) $\quad i=1,\ldots,m$

(ii) $p_i \sim$ Beta (a, b) $\quad i=1,\ldots,m$

This model is somewhat more flexible than the first model proposed, as the Negative Binomial has two parameters and can take on a wide range of shapes.

3. A Method for Model Checking

We need the following notations: (see Table 1).

Let i index the m countries in our group.

Let j index the 7 educational categories

Let p_i be the years of education to complete primary schooling in country i

Let s_j be the years of education to complete secondary schooling in country i

Let the number of years to complete higher education be four years across all countries (assumption as per Barro & Lee 1993)

Let f_{ij} be the number of people in category j in country i

Let n_i be the total number of people over 25 years old in country I

Table 1: Educational Categories

j	Barro & Lee Category	Description	J_{ij}	x_{ij}
1	NO25	%age of population over 25 whose highest educational attainment is no schooling	$\{0\}$	0
2	PRI25 - PRIC25	%age of population over 25 whose highest educational attainment is incomplete primary schooling	$[1, p_i)$	$\frac{1+p_i}{2}$
3	PRIC25	%age of population over 25 whose highest educational attainment is completion of primary schooling	$\{p_i\}$	p_i
4	SEC25 - SECC25	%age of population over 25 whose highest educational attainment is incomplete secondary schooling	(p_i, p_i+s_i)	$\frac{2p_i+s_i}{2}$
5	SECC25	%age of population over 25 whose highest educational attainment is completion of secondary schooling	$\{p_i+s_i\}$	p_i+s_i
6	HIGH25 – HIGHC25	%age of population over 25 whose highest educational attainment is incomplete higher schooling	(p_i+s_i, p_i+s_i+4)	p_i+s_i+2
7	HIGHC25	%age of population over 25 whose highest educational attainment is completion of higher schooling	$[p_i+s_i+4, \infty)$	p_i+s_i+4

Define

$$S_i = \sum_{j=1}^{7} f_{ij} x_{ij} ,$$

$$\bar{x}_i = \frac{S_i}{n_i},$$

$$n_T = \sum_{i=1}^{m} n_i ,$$

$$\bar{x} = \frac{\sum_{i=1}^{m} n_i \bar{x}_i}{n_T},$$

$$MSB = \frac{\sum_{i=1}^{m} n_i (\bar{x}_i - \bar{x})^2}{m-1},$$

$$d = n_T - \frac{\sum_{i=1}^{m} n_i^2}{n_T}.$$

Extending the basic ideas given in Ghosh and Lahiri (1997), we propose the method-of-moments estimators of different parameters in Model 1. Thus, equating MSB and $\bar{x}$ to their expected values under Model 1, we get estimates of α and β as follows:

$$\hat{\beta} = \frac{d\bar{x}}{(MSB - \bar{x})(m-1)} \text{ if } MSB > \bar{x} ,$$

$$\hat{\alpha} = \bar{x}\hat{\beta} .$$

With the estimated parameters of the Gamma distribution established, the λ_i for each country can be calculated, by finding λ at the mode of the posterior distribution. This is equivalent to maximizing the function below:

$$h_i(\lambda_i \mid f;\alpha,\beta) = (\alpha-1)\ln\lambda_i - \beta\lambda + \sum_{j=1}^{7} f_{ij} \ln(\pi_j(\lambda_i)),$$

where $\pi_j(\lambda_i)$ is the probability of an individual in country *i* (with $\lambda=\lambda_i$) falling into educational category *j*. Note that

$$\pi_j(\lambda) = P(X \in J_j \mid \lambda) = \sum_{u \in J_j} p(u;\lambda),$$

where $p(u; \lambda)$ is the probability function of F_2, in our case the Poisson distribution.

Once the α, β, and λ_i are available, expected frequencies for different education categories for each country are obtained as follows:

$$e_{ij} = n_i \pi_j(\lambda_i) \qquad (1)$$

In order to understand if the data fit our model, we will visually compare these expected frequencies with the observed frequencies. A statistical goodness-of-fit test like the Pearson's chi-square test cannot be directly applied to our situation because of the within

country dependence structure induced by Model 1. Moore (1978) provides an excellent review that one can use as a reference on the Chi-square tests.

Let us now discuss the parameter estimation for Model 2. Define

$$S_i^2 = \frac{\sum_{j=1}^{7} f_{ij}(x_{ij} - \bar{x}_i)^2}{n_i - 1}$$

$$SSB = \sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})^2$$

$$f = \sum_{i=1}^{m} n_i(1 - {}^{n_i}\!/_{n_T})\alpha_i^2$$

$$e = \sum_{i=1}^{m} (1 - {}^{n_i}\!/_{n_T})\alpha_i$$

$$\bar{\alpha} = \frac{\sum_{i=1}^{m} n_i\alpha_i}{n_T}$$

$$c = \frac{\bar{x}}{\bar{\alpha}}$$

$$SSB_\alpha = \sum_{i=1}^{m} n_i(\alpha_i - \bar{\alpha})^2$$

$$s = \frac{SSB - c^2 SSB_\alpha}{c(1+c)}$$

The method of moments estimators of α_I, a, and b are given by:

$$\hat{\alpha}_i = \left(\frac{S_i^2}{\bar{x}_i^2} - \frac{1}{\bar{x}_i}\right)^{-1}$$

$$\hat{a} = 1 + \frac{s+f}{s-e}$$

$$\hat{b} = (\hat{a} - 1)c$$

The method for checking Model 2 is similar to that of Model 1.

4. Data Analysis

We consider two groups of countries – those with high real GDP per capita (Group 1), and those with low level of real GDP per capita (Group 2). Tables 2 and 3 present summary statistics resulted from fitting Model 1.

Table 2: Results of Fitting Poisson-Gamma Model to Group 2

Country	*SHCODE*	S_i	$\bar{x}_i$	$\hat{\lambda}_i$		
Zaire	45	20581.43	1.869	1.379171	$\bar{x}$	1.922
Malawi	25	6557.476	2.657	1.925217		
Mali	26	1308.564	0.489	0.242334	MSB	2267.412
Uganda	44	8639.865	1.755	1.127716		
Niger	31	1055.688	0.4715	0.246461	d	72970.06
Burma	82	39130.54	2.581	2.277727		
Central_Af	9	1118.065	1.0855	0.710591	α	6.26176
Togo	42	2146.347	1.953	1.550435		
Mozambiq	30	4051.729	0.799	0.503003	$\hat{\beta}$	3.25
Gambia	16	205.03	0.707	0.568673		
Rwanda	33	3099.352	1.621	1.032403		
Sudan	39	10084.03	1.281	0.957087		
Ghana	17	11410.95	2.5035	2.009765		
Kenya	21	16211.36	2.6215	1.919269		
Zambia	46	7341.824	3.374	2.894097		
Liberia	23	1471.54	1.828	1.52358		
Sierra_Le	36	1450.076	1.0485	0.677154		
Haiti	57	5346.424	2.3035	1.90296		
Nepal	95	11016.57	1.645	0.651287		
Lesotho	22	2444.868	4.068	2.829659		

Table 3: Results of Fitting Poisson-Gamma Model to Group 1

Country	*SHCODE*	S_i	$\bar{x}_i$	$\hat{\lambda}_i$		
Canada	50	178268	10.365	10.41898	$\bar{x}$	9.72978
United States	66	1913964	12.0245	12.15493		
Hong Kong	84	29440.9	8.1735	8.177099	MSB	91402.9
Japan	90	759368	9.288	9.416541		
Singapore	100	9113.9	5.5845	5.299626	d	342899
Austria	107	37748.7	7.251	7.09921		
Belgium	108	54213.8	8.0115	8.120028	α	18.6941
Denmark	110	34002.1	9.7455	9.88482		
Finland	111	28652.6	8.5225	8.632031	$\hat{\beta}$	1.92133
France	112	255101	6.947	6.902312		
Iceland	116	1220.31	8.19	8.264738		
Italy	118	248320	6.481	6.266897		
Netherlands	121	84211.5	8.63	8.673017		
Norway	122	22365.9	8.0395	8.111732		
Sweden	126	53525.7	9.145	9.262227		
Switzerland	127	44334.1	9.465	9.529131		
United Kingdom	129	318040	8.3655	8.409812		
Australia	131	102632	9.7015	9.796191		

The λ_is were found using a quasi-Newton optimizer from the package SPLUS 3.2 Release 1 for Windows. Results were checked graphically in the neighborhood of the estimated solution. All were found to be unimodal and parabolic close to the estimated solution.

Table 2 shows the results of fitting the Poisson-Gamma model to the countries with low real GDP. Using (1), we find the corresponding expected frequencies, and comparing these to the observed frequencies, shows major differences between the two. The model almost always overestimates the number of adults who have incomplete primary schooling, while underestimating in all other categories. The differences are so large that calculation of the chi-squared test statistic is unnecessary. It is obvious that the model is not adequate for this data

Table 3 shows the results of fitting the Poisson-Gamma model to the upper income quintile countries. The fit is equally as poor as for the low-income countries, but in this case the model greatly over estimates the number of adults with incomplete secondary education, while underestimating those with complete secondary education.

Let us now turn our attention to Model 2. Table 4 provides summary statistics for group 2 countries.

Table 4: Results of Fitting Negative Binomial - Beta Model to Group 2

Country	*SHCODE*	S_i^2	$\bar{x}_i$	$\hat{\alpha}_i$		
Zaire	45	8.343103	1.869	0.539559	$\bar{x}$	1.921913
Malawi	25	11.25241	2.657	0.821328		
Mali	26	3.384144	0.489	0.082594	SSB	43080.83
Uganda	44	7.580515	1.755	0.528713		
Niger	31	2.777178	0.4715	0.096419	f	157061.9
Burma	82	13.36098	2.581	0.617957		
Central_Af	9	6.081522	1.0855	0.23585	e	22.94843
Togo	42	13.34043	1.953	0.334949		
Mozambiq	30	2.943041	0.799	0.297756	$\bar{\alpha}$	0.642348
Gambia	16	5.68074	0.707	0.100498		
Rwanda	33	7.187118	1.621	0.472078	c	2.992011
Sudan	39	7.888041	1.281	0.248365		
Ghana	17	14.9973	2.5035	0.50165	SSB_α	127501.1
Kenya	21	11.44334	2.6215	0.779006		
Zambia	46	12.86742	3.374	1.199133	s	-91955.2
Liberia	23	14.7167	1.828	0.259265		
Sierra_Le	36	8.012191	1.0485	0.157869	a	0.292151
Haiti	57	11.85375	2.3035	0.5556		
Nepal	95	4.695882	1.645	0.886965	b	**-2.11789**
Lesotho	22	5.180496	4.068	14.87522		

Note that the estimated value for *b* is negative, which is out of the allowable range for parameters of the Beta distribution. Furthermore, deleting Lesotho from the data set (because of its unusually high value of α_i actually worsens the problem – the estimate of *a* becomes negative as well).

The estimation for Group 1 countries is even more problematic, as some α_i's have very large estimates (greater than 100). The estimates of *a* and *b* are both negative for this group.

The problems encountered with the model fit indicate two likely problems:

1. The method of moments is inappropriate for this data, and should be replaced. One possible technique that could be used instead is simultaneously maximizing the posterior distribution on all parameters (a, b, $\alpha_i,\ldots,\alpha_m$).
2. This model is inappropriate for the data.

5. Conclusions

Neither of the models suggested here were totally successful in explaining the distribution of human capital (as expressed by highest educational attainment). However, they suggest that future models of this type will need to take into account the fact that schooling is more likely to be terminated at the completion of a category of schooling (e.g. primary, secondary, and higher education) than during a category. While this is a rather logical observation from the real world, it is clear that the simple probability models discussed above do not give results of this type. It is necessary to either extend the model to account for this "discontinuity" in the data, or "smooth" the data by ignoring the data about completion of education (ie PRIC25, SECC25, and HIGHC25) and modelling only NO25, PRI25, SEC25 and HIGH25.

Acknowledgements
The research has been partially supported by the Marsden Grant # 96-UOA-SOC-0018 of the Royal Society of New Zealand. The author thanks Professor P. Lahiri of the University of Nebraska-Lincoln, U.S.A., a co-investigator of the Marsden project, for providing help in statistical analysis and noting the need for extending goodness-of-fit tests in the non iid situations.

References

Bandyopadhyay, D. (1993), "Distribution of Human Capital, Income Inequality and the Rate of Growth," *Ph.D thesis, The University of Minnesota.*

----------, (1997), "Distribution of Human Capital and Economic Growth," *The University of Auckland, Department of Economics Working Paper Series*, No. 173.

Banerjee, A. V. and A. F. Newman, (1994), "Poverty, Incentives and Development," *American Economic Review*, Vol, 84 (2), 211-15.

Barro, R. J. and J.-W. Lee, (1993) "International comparisons of Educational Attainment," Journal of Monetary Economics 32, 363-394.

Barro, R. J. and J.-W. Lee, (1997) "International Measures of Schooling Years and Schooling Quality," American Economic Review 86(2), 218-223.

Chari, V. V. And H. Hopenhayn, (1991) "Vintage Human Capital, Growth, and the Diffusion of New Technology," Journal of Political Economy 99(6), 1142-1165.

Galor, O. and J. Zeira, (1993), "Income Distribution and Macroeconomics," *Review of Economic Studies*, 60, 35-52.

Galor, O. and D. Tsiddon, (1996) Income Distribution and Growth: The Kuznets Hypothesis Revisited.. Economica. Vol. 63 (250). p S103-17.

Ghosh, M., and P. Lahiri (1987), "Robust empirical Bayes estimation of means from stratified samples," Journal of the American Statistical Association, 82, 400, 1153-1162.

Kaneko, M., (1986), *The Educational Composition of the World's Population: A database.* Washington, DC, The World Bank, Education and Training Department. Report No. EDT 29.

Krueger, A. B., (1993), "How Computers Have Changed the Wage Structure: Evidence from Microdata, 1984-1989," *Quarterly Journal of Economics*, 108, 33-60.

Krueger, A. O., (1968) "Factor Endowments and Per Capita Income Differences among countries," Economic Journal LXXVIII, 641-59.

Lau, L. J. et-al, (1993) "Education and Economic Growth: Some Cross-Sectional Evidence from Brazil," Journal of Development Economics 41(1), 45-70.

Lucas, R. E., (1988) "On the Mechanics and Economic Development," Journal of Monetary Economics 22, 3-42.

Mankiw, N. G., D. Romer, and D. N. Weil, (1992) "A Contribution to the Empirics of Economic Growth," Quarterly Journal of Economics 107, 407-437.

Mincer, J., (1991), "Human Capital, Technology, and the Wage Structure: What Do Time Series Show?" NBER Working Paper No. 3581.

Moore, D.S. (1978), "Chi-square tests," in Studies of Statistics, R.V. Hogg Ed., Studies in Math., 19, Math. Asso. Amer.

Nehru, V.; Swanson, E. and A. Dubey, (1995), "A New Database on Human Capital Stock in Developing and Industrial Countries: Sources, Methodology, and Results," *Journal of Development Economics*. Vol 46 (2), 379-401.

Psacharopoulos, G. and A. M. Ariagada,, (1986), *The Educational Attainment of the Labour Force: An International Comparison*, Report No. EDT 38 (Education and Training Department, The World Bank, Washington, DC).

Psacharopoulos, G. and M. Arriagada, (1986) "The Educational Composition of the Labour Force: An international comparison," International Labour review 125(5).

Romer, P. M., (1990) "Endogenous Technological Change," Journal of Political Economy 98(5), s71-s99.

Romer, P. M., (1989) "Human Capital and Growth: Theory and Evidence," NBER Working paper No 3173.

Romer, P. M., (1986) "Increasing Returns and Long-Run Growth," Journal of Political Economy, 94(5), 1002-1037.

Solow, R. M., (1956) "A Contribution to the theory of Economic Growth," Quarterly Journal Economics LXX, 65-94.

Togo, K., (1996) "The Distribution of Human Capital and Economic Growth: Application of the Japanese Experience between 1868 and 1990," PHD thesis, Yale University.