PN-ABC-724

RESEARCH TRIANGLE INSTITUTE

A New Approach To The Estimation of Enrolment Transition Rates: Estimating Repeater and Dropout Rates with Minimal Data Availability

Luis Crouch
Research Triangle Institute
BRIDGES Project
March 1989

1036
1825

Work on this paper was funded by the Basic Research and Implementation for DevelopinG Education Systems (BRIDGES) project, which is sponsored by the U.S. Agency for International Development, Office of Education, under Cooperative Agreement No. DPE-5824-A-5076.

POST OFFICE BOX 12194 RESEARCH TRIANGLE PARK, NORTH CAROLINA 27709-2194

A Simplified Linear Programming Approach to the Estimation of Enrolment Transition Rates: Estimating Rates with Minimal Data Availability

Luis Crouch
Research Triangle Institute
BRIDGES Project
March 1989

Introduction: The Problem

Knowing rates of repetition, dropping out, and entrance into first grade is of great importance to educational planners and policy analysts in developing countries. Knowledge of these rates is essential to the accurate forecasting of enrolments and to the analysis of the internal efficiency of school systems. Standard methodological manuals explain how to calculate all these rates. (See, for example, Thonstad.) Unfortunately, the traditional methods of estimating these rates are fraught with serious problems. This paper suggests an alternative to the traditional method. The proposed alternative requires data available in all countries, and can be quickly computed using a microcomputer.

There are two fundamental problems with estimating repetition and dropout rates according to traditional methods. The first is the simplest: many demographically important countries (e.g., Egypt, Pakistan) simply do not report repeater or dropout data on any systematic basis. The second problem is that in many countries where the numbers of repeaters are reported, the numbers are consistently undercounted, and the amount of bias is usually unknown. Since the repetition rate is estimated as the number of students repeating a grade at time t divided by the enrolment in that grade at time t-1, the repetition rate will evidently be underestimated. Since the dropout rate is usually estimated as a residual, this rate will be overestimated. The rate of entrance into the first grade is also frequently estimated as a residual, and is thus also overestimaded. The result will be that both policy makers and planners will be misled. Policy makers will tend to think that repetition is a less serious problem than it really is, and concentrate their policies on preventing dropping out. At the same time, they will be misled into thinking that the percentage of children entering school out of every population cohort is much higher than it really is, and they will tend to think the problem of access to schooling is close to being solved. Planners, on the other hand, if they use these rates to forecast enrolment, will tend to make forecasts that are too low, since

they assume repetition rates to be lower, and dropping out rates to be higher than they really are.

A natural question then arises: could the rates of repetition and dropping out be estimated without reference to the absent or mistaken counts of repeaters? If we can assume that actual enrolment data are either better or at least consistently biased, are there some mathematical or statistical techniques that could be used which depend only on information about actual enrolments? Under these assumptions, a method of estimating repetition and dropping out rates that depended only on enrolment data might yield better estimates of these rates. Several logical alternatives immediately come to mind.

Extant Alternative Solutions

A first alternative is to think about solving systems of equations. Let's take first grade enrolment as an example, where the problem is that of estimating the first grade repetition rate and the rate of entrance into the school system. We know that the equation for first grade enrolment is

$$E_{1,t}=r_{1,t-1}E_{1,t-1} + e_tP_t$$

where $E_{1,t}$ is enrolment in first grade at time t, $r_{1,t-1}$ is the first grade repetition rate at time t-1, e_t is the rate of entrance of new students from the population of school entrance age into the first grade, and finally P_t is the population of school entrance age at time t. Now, under the assumption that the repetition and entrance rates do not change radically from year to year, and if we had three years of data on enrolment and two years of data on population of entrance age, we would have two equations in two unknowns,

$$E_{1,t}=r_1E_{1,t-1} + eP_t$$

$$E_{1,t-1}=r_1E_{1,t-2} + eP_{t-1}$$

where the unknowns are the repetition rate and the entrance rate. Note that we have removed the time subscript from the rates to denote the fact that we are assuming the rates do not change from t-2 to t-1. The problem with this method is that the equations are so similar in slope that the solution tends to give results that are mathematically correct, but make no sense as enrolment transition rates. Take as an example the following data for first grade enrolment in two recent years in the Dominican Republic (all schools, both sexes):

Year	E_t	E_{t-1}	P_t
1	310277	308478	159873
2	316605	310277	161464

The first column represents the right hand side of the equation above. The second column, representing last year's enrolment, would be multiplied by a repetition rate. The last column, representing this year's population, would be multiplied by the entrance rate. The solution to the two implicit equations in the table above, unfortunately, turns out to give a repetition rate of -2.54, and an entrance rate of 6.86. Obviously, this method can not be offered as a general method. There are two potential problems with this method. First, it may seem that the key problem is the assumption that the rates do not change between the two years. As we will see later, however, this is actually a reasonable approximation. A second, more serious problem is that, given the assumption of exact equality of the rates in the two years, forcing exact equality in the relations determining enrolment, when the relations are so similar, yields estimated rates that can easily be in an "unreasonable" range. The advantage of the simultaneous equations approach, evidently, is that very few data are required: only three consecutive years of enrolment data and two consecutive years of population data.

A second alternative that seems intuitively appealing is to take longer time series and apply a "least-squares error" criterion to fitting the crucial rates. We would have a long time series of data on E_t, which would assume the role of "dependent" variable, and on E_{t-1} and P_t, which would assume the role of "independent" variables. We would then estimate our repetition and entrance rates as those which minimize the differences between the observed E_t and the E_t predicted with the estimated rates. This procedure would have the advantage of not requiring an exact fit with the observed data (i.e., it is rather more of a "statistical" rather than a "mathematical" approach), and so hopefully the estimated rates would make more sense. Unfortunately, it requires assuming that the rates do not change over a period much longer than two years, which is a much more restrictive assumption. Furthermore, the two "independent" variables (E_{t-1} and P_t) are highly correlated with each other, resulting in what statisticians term "multicollinearity", which means that the estimated rates will be quite unreliable from a statistical point of view. (As well as possibly meaningless from an educational point of view.) Finally, estimation of the transition rates using this method could be plagued by autocorrelation.

In response to these problems, educational planning researchers have been searching for alternative methods to

estimating these rates. Schiefelbein, for example, has developed various approaches, although (to varying degrees) those methods require more extensive data than the method we propose here.

A New, Simplified Approach

The approach proposed in this paper attempts to deal with the problems mentioned above. It does not require exact equality in the relations, as in the simultaneous equation option. It does not require a long time series of data, as in the least-squares option; nor does it require the shaky assumption that transition rates do not change over the longer period. It does not require data on enrolment by age and grade, which are not available in many countries. The method proposed has minimal data requirements, and the assumptions can be made as restrictive or as relaxed as the user feels comfortable with. Only three years of consecutive enrolment data, and two years of data on population of school entrance age are required.

The method has the advantage of mathematically maximizing the fit between the enrolment pattern produced by the estimated rates and the historical pattern in the three years of enrolment data. There are two disadvantages, however. The first is a methodological disadvantage: it may occasionally produce rates that seem unstable from grade to grade. (What this means will become apparent when we present an example.) Thus, the knowledge produced by this method may need to be tempered with knowledge derived from more traditional methods. In this sense, this method is offered not as an alternative to any others, but as a complement. The second disadvantage is that it is impossible to apply by hand: it requires access to a microcomputer and to linear programming software. Fortunately, most planning and policy offices in Ministries of Education in developing countries now have access to microcomputer hardware, and getting linear programming software is not at all expensive.

What does the proposed method consist of? We saw above a description of the problem of estimating enrolment transition rates as a simultaneous equations problem. But we also saw that the straightforward use of simultaneous equation solutions led to results that are mathematically correct but make no sense from an educational point of view. What if we can take a modified simultaneous equation approach? Part of the problem, we saw, was that the "exactness" of the method was causing problems. What if one can relax the "exactness" aspect of simultaneous equations, and yet keep the "low data requirement" aspect? What if we could find a less "exact" method, e.g., one that could allow for inequalities? Furthermore, what if this method allowed one to incorporate restrictions that prevented results that might be mathematically correct but made no educational sense? One available method is to use the simplex algorithm of linear

programming as a solution tool to maximize the fit between observed enrolment and the enrolment "predicted" by a set of estimated rates.

We pose the problem thus: we need to determine repetition rates for grades 1 through 6, an entrance rate, and dropout rates for grades 1 through 5. Thus, we have twelve unknowns. Are there twelve linear mathematical relations between these unknowns that we could use? If we assume that the rates do not change between a given two sets of two consecutive years, then we can have one fundamental enrolment equation, for each of six grades, for each of two years: twelve relations. For any given year, and for any grade other than first grade (we have seen the first grade equation above), this equation is

$$E_{g,t} = r_{g,t-1}E_{g,t-1} + (1-r_{g-1,t-1}-d_{g-1,t-1})E_{g-t,t-1}$$

Let us suppose, however, that we do not insist on perfect equality as in the relation above, and at the same time assume that the rates do not change from t-2 to t-1. Let us call the right hand side of the equation above, with the estimated transition rates, the "predicted" enrolment, E^*. Thus, we would have one equation

$$E^*_{g,t} = r^*_g E_{g,t-1} + (1-r^*_{g-1}-d^*_{g-1})E_{g-t,t-1}$$

for each grade in each of two years. Note that we have removed the time subscript from the rates, to indicate that they hold for both years. Now, the trick is to estimate transition rates so as to minimize the deviation the "predicted" values and the real values. Since there is no a priori reason to prefer a squared error criterion to an absolute error criterion, and since the latter is easier to handle within the context of linar programming, we minimize the sum of the absolute errors:

$$\text{Min } \Sigma_{g,t} \text{ abs } (E^*_{g,t} - E_{g,t}) \text{ over } r^*_g, d^*_g, \text{ and } e^*,$$

which is equivalent to

$$\text{Min } \Sigma_{g,t} E^*_{g,t} \text{ over } r^*_g, d^*_g, \text{ and } e^*,$$

subject to

$$E^*_{g,t} > E_{g,t}.$$

In addition, the following constraints need to be satisfied in order to define the relationship between the rates to be estimated and the predicted enrolment:

$$E^*_{1,t} = r^*_1 E_{1,t-1} + \quad e^* P_t$$

$$E^*_{g,t} = r^*_g E_{g,t-1} + (1-r^*_{g-1}-d^*_{g-1})E_{g-t,t-1}$$

Furthermore, arbitrary but reasonable limits can be placed on various rates with nearly infinite flexibility and very explicitly, as long as these limits can be expressed linearly. For example, one might constrain the estimated entrance rate e^* so that $e^* < 1.0$, or some other reasonable upper bound. Note we have removed the time subscripts from the rates in the equations above, to signify that the same rates apply for both time periods under consideration.

All this can easily and quickly be solved with any linear programming package on a small microcomputer (e.g., **What's Best**), though it would be a true nightmare to try to do with a hand calculator. It could possibly be solved once with a hand calculator, but this would not be enough to allow the planner or policy analyst to "experiment" with the method. And it is this experimentation which is essential to the process of planning and policy analysis. With a microcomputer it takes literally minutes to set up and solve.

An Example

Suppose we have only the following data (national level data from the Dominican Republic, boys and girls combined):

	Population	Enrolment					
		Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6
Year 1	159873	308478	173528	140861	118826	92252	74646
Year 2	161464	310277	177423	145347	123951	97173	79687
Year 3	163055	316605	178724	149076	126908	101395	82395

The problem then is, using only these data, to estimate the rates of transition between all the grades and between the population and the first grade.

The initial results with the constraints as listed above, are:

	Estimated transition rates		
	Entrance	Repetition	Dropout
Grade 1	1.0	0.49	0
Grade 2	---	0.12	0.13
Grade 3	---	0.12	0
Grade 4	---	0	0.21
Grade 5	---	0.03	0.10
Grade 6	---	0	---

Clearly there is a problem in that, as has been mentioned before, the rates are somewhat unstable from grade to grade. For example, it is unlikely that the repetition rates for grades 4 and 6 are really 0, or that the dropout rates for grades 1 and 3 are similarly 0. One of the advantages of the linear programming method as an approach is that the mechanical aspects of the methodology can be tempered with additional empirical information coming from many disparate sources. Suppose we have good reason to suspect that the rates for any given grade are not likely to differ by more than 50% from those of the next grade up or down, and that the average rates across all grades must be less than 30% but greater than 5%. Formally, this means adding the following constraining relations to the mathematical statement of the problem:

$r_g > 0.5\ r_{g-1}$

$r_g < 1.5\ r_{g-1}$

$d_g > 0.5\ d_{g-1}$

$d_g < 1.5\ d_{g-1}$

$\Sigma_g\ r_g < 0.3 * 6$

$\Sigma_g\ r_g > 0.05 * 6$

$\Sigma_g\ d_g < 0.3 * 6$

$\Sigma_g\ d_g > 0.05 * 6$

With these modifications, the method provides the following estimates of the transition rates, which seem much more reasonable, and yet are produced without using any official repetition data:

Estimated (constrained) transition rates

	Entrance	Repetition	Dropout
Grade 1	1.0	0.49	0.07
Grade 2	---	0.25	0.05
Grade 3	---	0.16	0.06
Grade 4	---	0.12	0.11
Grade 5	---	0.06	0.10
Grade 6	---	0.03	---

Adding any assumptions and prior information can thus temper the mechanical results produced by the method. Fortunately, the assumptions can be expressed in a very clear and explicit manner, and the effects of the assumptions on the estimated rates are easy to test. Note, for example, that the estimated repetition rate for the first grade has barely changed, and that repetition rates as estimated here are consistenly higher than dropout rates, even with the new assumptions. The method allows the analyst to rapidly test the effects of many sets of reasonable restrictions. If certain results are robust to the various sets of reasonable assumptions (for example, if the first grade repetition rate remains between 0.4 and 0.6 under all reasonable scenarios), then it is safe to conclude that those results are likely to be a true reflection of the underlying rate structure.

Conclusion

The method presented in this paper has the capability of offering the analyst an alternative to traditional methods of estimating repetition and dropout rates. This alternative method relies only on widely available data, and can explicitly incorporate any prior knowledge about the rates. Evidently, it is not an exact method. But, in combination with other available methods, it can be used to try to get a closer fix on what the true repetition and dropout rates are in many developing countries.

REFERENCES

Schiefelbein, Ernesto, and M.C. Grossi. 1980. **Statistical Report on Repetition in Latin America.** UNESCO, Division of Statistics on Education, Office of Statistics.

Thonstad, Tore. 1980. **Analysing and Projecting School Enrolment in Developing Countries: A Manual of Methodology.** UNESCO, Statistical Reports and Studies Series No. 24.

What's Best, General Optimization, Inc., Chicago.