Understanding Student Growth Percentiles

Introduction. Standardized test scores have long been used to characterize educational achievement at a point in time.

More recently the focus has shifted to longitudinal methods that attempt to quantify the "growth" of educational achievement for individual students over multiple years.

The Colorado or Student Growth Percentile model is one of a number of proposed approaches to this problem.

So What Exactly is "Growth"? The SGP model gets around the lack of a vertical scale by redefining the meaning of "growth". In the words of SGP's principal architect Damien Bettebener,

> *A students growth percentile describes how typical a students growth is by examining his/her current achievement relative to his/her academic peers - those students beginning at the same place. That is, a student growth percentile examines the current achievement of a student relative to other students who have, in the past, walked the same achievement path.*

For a more precise definition, Bettebener falls back to a description of how it is computed:

> *Quantile regression is used to establish curvi-linear functional relationships between the cohorts prior scores and the cohorts current scores. Specifically, for each grade by subject cohort, quantile regression is used to establish 100 (1 for each percentile) curvi-linear functional relationships between the students grade 3, grade 4, grade 5, and grade 6 prior scores and their grade 7 scores. 4 The result of these 100 separate analyses is a single coefficient matrix that can be employed as a look-up table relating prior student achievement to current achievement for each percentile.*

The Massachusetts Board of Elementary and Secondary Education (MBESE) announced the implementation of growth reporting with the statement:

> *For over a decade, MCAS scaled scores and performance levels have answered the question, "How much has this student achieved compared to the state's grade-level learning standards?" The new growth score, called a Student Growth Percentile (SGP), answers the question, "How much did a student grow over the previous year compared to his or her academic peers?*

Once again, we have only a vague definition of "growth". It would be more accurate to say that it answers the question "Where does the student's current score rank among the scores of students who had similar scores in previous years?".

NURE 2014: The "Black Box" Approach. Lacking a mathematically precise definition of "growth", last summer (and up to recently, this summer) we chose to treat the SGP as a "black box": test score history in, SGP out.

Simulation is common approach to analyzing systems of this type. We build a large number of artificial inputs with various characteristics, and observe the growth percentiles the black box produces.

The simulation itself is pretty straightforward (we have the mechanism in place with driver.py and the R code we have developed). The main difficulty is that there are a lot variables to manipulate and the simulations take a long time to run.

NURE 2015: an Analytic Approach. Analytic methods are preferable to simulation if they are available. If we recognize that the collective purpose of the 99 quantile regression models is to produce an estimate of the conditional cumulative distribution function (CDF) of the current year's score given the logitudinal history, we can give a mathematically concise definition of "growth":

Definition 1 (growth score)**.** *Given a historical record of n standardized test scores from previous years $x_1, x_2, \ldots, x_n$, and a random variable Y representing the current test score with observed value y, the* ***growth score*** *$S(y)$ associated with y is:*

$$S(y) = 100 \cdot P(Y \leq y | x_1, x_2, \ldots, x_n)$$

In other words, $S(y)$ is 100 times the conditional cumulative distribution function (CDF) of Y given $x_1, x_2, \ldots, x_n$ evaluated at the observed value y.

The importance of having this definition is that for the first time, we can say what the **true** theoretical value of the growth score is for an individual student, without having to compute the SGP for an entire cohort of students (70,000).

I ran an artificial example using a "test" consisting of a few dichotomous items and the results agree with the above definition with a couple of caveats:

- To get the same result as SGP, you have to replace $P(Y \leq y)$ with $P(Y < y)$.
- SGP discretizes the conditional CDF of test scores into percentiles from 1 to 99 (or 0 to 100 if you request it).

For a continuous distribution, neither of these differences matter because the probability of exact equality $P(Y = y)$ is always zero. This is not true for a discrete distribution, which is what we have.

Application to Teacher Evaluation. If we think of a simplified teacher evaluation system as classifying individual teachers as "good" or "bad", the power of the system would be defined as its ability to separate the two kinds of teachers.

While this is an oversimplification, many actual evaluation systems classify teachers into four categories ("quartiles") or five categories ("quintiles"). There is empirical evidence that these classifications are not very stable.

A fair teacher evaluation system must have a high power value (at least .8). As the power decreases, the system looks more and more like Demming's "Red Bead Experiment".

When Professor Stuart Yeh describes VAM-based evaluation systems as essentially a coin toss, he is saying that their power is very low.

At this point in the debate, there are a fair number of research articles that discuss VAM, but very few that discuss SGP.

So what can we say we have accomplished? I think having a concise definition of growth that we can show to be equivalent to SGP is a contribution in its own right, and shows a lot of promise as a technique for examining the behavior of proposed teacher evaluation systems. regardless of whether we have time to explore this or not.

I think there is enough time for us to firm up the relationship of the SGP and our concise definition of growth with some examples. These could be artificial like my binomial "test", or based on simulated MCAS scores using our R code.