Can flexible non-linear modeling tell us anything new about educational productivity?

Bruce D. Baker

Department of Teaching and Leadership, The University of Kansas, Lawrence, KS 66045, USA

Received 28 November 1998; accepted 24 March 1999

Abstract

The objective of this study is to test, under relatively simple circumstances, whether flexible non-linear models — including neural networks and genetic algorithms — can reveal otherwise unexpected patterns of relationship in typical school productivity data. Further, it is my objective to identify useful methods by which “questions raised” by flexible modeling can be explored with respect to our theoretical understandings of educational productivity. This study applies three types of algorithm — Backpropagation, Generalized Regression Neural Networks (GRNN) and Group Method of Data Handling (GMDH) — alongside linear regression modeling to school-level data on 183 elementary schools. The study finds that flexible modeling does raise unique questions in the form of identifiable non-linear relationships that go otherwise unnoticed when applying conventional methods. © 2001 Elsevier Science Ltd. All rights reserved.

JEL classification: I21

Keywords: Neural networks; Functional form

1. Introduction

Production function applications to educational research have gradually, but steadily, evolved since the Coleman Report of 1966. Among the current trends are increased emphasis on student-level data (Goldhaber & Brewer, 1996), school-level data (Harter, 1999; Murnane & Levy, 1996), greater understanding of the hierarchical design of our system of schooling (Kaplan & Elliott, 1997),¹ and the relevance of the structural nature of direct and indirect relationships within that system (Kaplan & Elliott, 1997). In addition, substantial emphasis has been placed on identifying more useful measures of the outcome — “educational productivity”. Progress in outcome measurement has, however, led to a divergence rather than a convergence of philosophies, with current preferences ranging from economic impacts on the labor market and earnings (Betts, 1996; Card & Krueger, 1996) to more basal school achievement measures such as minimal concept mastery (Harter, 1999).

Despite the apparent substantive progress made with respect to conceptual and methodological concerns, a few basic rules continue to govern production function methodologies. First, production function studies are typically performed within the narrow confines of formal deductive hypothesis testing. That is, the researcher begins with a question along the lines of — “Are school-level instructional expenditures per pupil related to student achievement outcomes?” Next, the researcher establishes his/her hypothesis, based on prior research and theoretical assumptions regarding the expected outcomes, and constructs a statistical model for testing the hypothesis.

Although a well-understood and generally accepted
paradigm, this purely deductive approach presents certain potential difficulties to the researcher. For one, this approach requires that the researcher has or finds some prior knowledge as to how the system in question works. This knowledge may ultimately be rooted in anything from valid theoretical constructs to personal or political biases, the latter of these problems becoming more prevalent when dealing with politically heated issues like educational productivity.

Problems associated with a priori understanding of the system are confounded when applied to the development and application of a statistical model for hypothesis testing. Typically, the production function is expressed as follows:

\[ f(Q, X, S) = 0, \]

such that outcomes, \( Q \), are a function of schooling inputs, \( X \), and non-school inputs, \( S \). This function is most often analyzed in the form of a linear regression equation:

\[ Q_{ij} = \beta X_{ij} + \gamma S_{ij} + \varepsilon_{ij}, \]

where \( Q_{ij} \) is the outcome of student \( i \) in school \( j \), \( X_{ij} \) are the schooling inputs to that student, \( S_{ij} \) is a vector of non-schooling inputs and \( \varepsilon \) is a stochastic error term.

Linear regression applications to production function modeling and related estimation procedures are limited in a variety of ways. These limitations include, but are not limited to, difficulties with the selection of model parameters. The usefulness of linear regression models lies in our ability to interpret individual regression coefficients, their statistical significance and respective magnitudes. More meaningful models tend to be those that are parsimonious, addressing necessarily narrow questions as exemplified by Goldhaber and Brewer (1996) in their study of the effects of teacher characteristics on student performance outcomes. While such studies provide valuable insights with respect to the question at hand, the necessity to repeatedly narrow research questions to this degree increases the probability that education researchers and economists may miss potentially important questions.

Other studies have relied on data dumping of massive numbers of potential inputs, neglecting the effects of multicollinearity on both the magnitude and significance of the regression coefficients of interest. Harter (1999), for example, separately includes salaries and benefits for each category of personnel in Texas schools, eventually concluding none of them to be significant, but finding more obscure measures such as salary supplements to be positively related to performance and substitute pay to be negatively related to performance. While data dumping may in some ways serve as a useful preliminary step in such complex analyses, it is unlikely to yield clear, definitive or even useful results in linear regression modeling. In addition, tools such as step-wise regression for selecting more parsimonious linear regression models from among the various predictors are generally inadequate.

Selection of functional form is similarly problematic in regression modeling. Linear regression modeling, by definition, seeks to identify linear relationships between specified input and outcome measures. That is, relationships are assessed on the extent to which unit increases in \( X \) are constantly related to unit increases in \( Y \). Hanushek (1996, p. 55), for example, discusses 90 studies which collectively generate 377 attempts to estimate a linear, or some highly restricted variant, relationship between teacher–pupil ratios and/or teacher education and student performance. Hanushek concludes that no systematic (linear) relationship exists.

We would perhaps be wise to consider the possibility, if not the probability, that within matrices of data on schooling productivity, there are actually some non-linear relationships that are “tighter” than some linear relationships. These relationships, where their curvilinear nature substantially violates assumptions of linearity, may go unrecognized or their magnitude underestimated when using linear methods (Cohn & Geske, 1990, p. 166). The only way to identify these relationships via conventional methods is to know or at least expect in advance that they exist and integrate them into econometric models as higher-order terms or alternative functional forms.

A common a priori assumption of non-linearity rooted in economic theory is that of diminishing returns. As noted by Betts (1996, p. 163), “the education production function, like all well-behaved production functions

---

2 I choose the term dumping here where others might use data mining. The intent is not to choose a more derogatory term, but to use a term that emphasizes the distinct difference between this method and data “mining” methods discussed later. Data dumping, as it is used herein, refers to attempting to include all possible variables in a single model whereas data mining is used to describe a process of sifting through all possible variables to find potential relationships that may be modeled.


4 Similarly, in structural models such as those applied by Kaplan and Elliott (1997), each direct effect in the structural equation model is represented as a linear relationship. Although combinations of direct and indirect effects may yield non-linearities, structural equation models do not explicitly allow for sets of non-linear direct effects. It is presumed that these relationships could be accommodated by either data re-scaling (log–log relationships) or the inclusion of higher-order terms (squared, cubed, etc.).

5 For example, linear relationship between logged (ln) terms and other functional forms to be discussed later.

6 In terms of R-squared if fitted with a curve.
[emphasis added] is subject to diminishing returns. This behavior is generally well captured by applying a log-log specification of wages relative to per pupil spending. Others have replaced the log of spending with a quadratic function, achieving a similar interpretation (Johnson & Stafford, 1973). More recently, Figlio (1999) questioned the effectiveness of highly restrictive specifications of functional form for estimating education production functions, noting in particular the usefulness of more flexible estimation procedures.

Betts’ choice of the phrase “well-behaved” is indicative of the standard mindset with which we approach production function modeling. This common econometric phrase suggests that our primary objective as a researcher is to determine the extent to which reality or data generated by the underlying processes of our reality “behaves” according to the mathematical specification of our mental model of that process. An underlying presumption being that if the data fail to conform to our model, that there is either some flaw in the data or the system, rather than a flaw in our mental model. In light of this perspective, an appropriate re-framing of Hanushek’s (1996) conclusion might be: “We have yet to generate statistical findings to support that the relationship between teacher–pupil ratios and/or teacher education and student performance conforms to our mental model for that relationship.”

Complementary inductive methods and analytical tools do exist — some of which can specifically provide support in the areas of parameter selection and identification of potential non-linear forms. The methods demonstrated in this study fall under the broad analytic umbrella of Data Mining. Data mining is the process of exploring the available data for patterns and relationships (Lemke, 1997). Data mining activities range from the visual exploration of bivariate scatterplots, often done as a preliminary to formal econometric modeling, to the use of iterative pattern learning algorithms or neural networks to search for potential relationships in data sets. While it is presumed that the development of most econometric models involves a great deal of inductive tinkering by the researcher, it is also presumed that the researcher cannot efficiently explore all possibilities or conceptualize the plethora of non-linear relationships that may exist. In addition, the human researcher brings with him/her the baggage of personal and political predisposition as to what the data should say. Thus, this study explores the use of flexible non-linear modeling, including neural networks, as a supplement to the typical preliminary activities of induction, and as a complement to conventional deductive production function analysis.

Broadly speaking, neural networks are iterative pattern learning algorithms modeled after the physiology of human cognitive processes. Unfortunately, the term “neural network” is also frequently misused as an overarching classification encompassing other types of algorithms, including genetic algorithms, that achieve similar ends, but by different means. This study applies both neural and genetic algorithms and refers to them collectively as flexible non-linear models.

Applied to econometrics, flexible non-linear models are free of a priori assumptions of functional form, deriving deterministic equations from available data, selecting predictors that best serve the modeling objective — prediction accuracy. Cross-sectional predictive and time-series forecasting accuracy of flexible non-linear models has been validated in the fields of medicine (Buchman, Kubos, Seidler & Siegforth, 1994), real estate valuation (Worzala, Lenk & Silva, 1995), bankruptcy assessment (Odom & Sharda, 1994) and forecasting education spending (Baker & Richards, 2000). Others have noted the potential usefulness of neural networks for exploring data in social science research (Liao, 1992).

The objective of this study is to test whether flexible non-linear models can reveal otherwise unexpected patterns of relationship in typical school productivity data. This study builds on the work of Figlio (1999) by combining flexible functional form with inductive estimation algorithms to provide a more sensitive estimation of potential relationships in the given data set. The ultimate goal of this exercise is to identify useful methods and develop a framework by which “questions raised” by flexible modeling can be explored with respect to our theoretical understandings of educational productivity. This study applies three types of flexible estimation procedure, alongside linear regression modeling, to school-level data on 183 elementary schools.

2. Methods

2.1. Data

All data were acquired through the Vermont Department of Education (1998) and are contained in the annual release of the Vermont School Reports. Although individual student-level data were not available, this study
is in keeping with current convention in that all analyses are applied to school-, rather than district-level data. The study focuses specifically on 183 elementary schools. The author recognizes the data set as relatively crude and very small, but similar in nature to data commonly used in productivity analysis. Recall that the primary objective is not to derive meaningful policy implications for the state of Vermont, but to explore the potential usefulness of a new methodology.

The outcome measure of interest is the average percentage of students passing mathematics competencies in Grade 4 on the New Standards Assessments. Potential input measures include the average enrolment per grade level as an index of school size, the length of school day in hours, the length of school year in days, the average class size, the portion of students receiving special education services, the student-to-computer ratios for the school and a set of community socio-economic variables expected to display some collinearity, including the percentage of families below the poverty line, the percentage of parents with postsecondary education and a combined index of property wealth per pupil and median income of the head of household (detailed descriptions provided in Appendix A).

2.2. Model comparisons

Three flexible algorithms were compared with linear regression modeling. These include backpropagation, generalized regression neural networks (GRNN)\textsuperscript{10} and group method of data handling (GMDH).\textsuperscript{11} Baker and Richards (2000) applied the same architectures, using the same software,\textsuperscript{12} to forecasting educational spending and present a more thorough description of the model estimation procedures than will be given here. In short, however, the key to neural networks is that the ultimate objective is prediction accuracy, or the ability to predict accurately the outcome measures of non-sample data, having been trained (estimated) on sample data. This objective yields different results from simply trying to achieve the greatest fit to sample data. It has been shown that while increasing model complexity yields asymptotically tighter fits to sample data, the ability to predict non-sample data begins to erode beyond a given level of complexity.\textsuperscript{13} As a result, deterministic models generated by neural estimation can be considered generalizable.

To achieve the ability to predict accurately non-sample data (to be referred to as a production set), neural networks typically estimate deterministic non-linear equations using sample data subdivided into classes — training set data and test set data. By a process of randomly drawing weights or coefficients, a neural network will iteratively reduce error on the training set, while occasionally, at defined intervals (200 passes or epochs), checking its ability to predict the outcomes of the test set. When several (20) attempts have been made to improve test set prediction error with no gain, the network stops, saving the weights that achieved the best test set prediction. It is presumed, and often holds, that this simulation will result in most accurate prediction of non-sample data.

Table 1 displays the algorithms selected and the training set, test set and production set structures for the analyses. The same production set data \((n=20)\) were extracted for all comparisons of prediction accuracy. For the linear regression model, ordinary least-squares (OLS) estimation was applied to 163 elementary schools. For the backpropagation model, iterative estimation was used giving a training set of 131 schools and test set of 32 (20\% as recommended by Ward Systems Group, 1995, p. 101). Unlike the iterative convergent method of applying weights used in backpropagation, GRNN involves generating pools of equations and selecting those measured to be most fit, determined by test set prediction error. This approach is known as a genetic algorithm. In this case, 20 generations without prediction improvement is the stop criterion. Finally, GMDH involves identifying a best predicting polynomial equation by breeding and hybridizing polynomials of increasing complexity. The main difference between GMDH and GRNN estimation, given the software selected for this study, is that the GMDH-type\textsuperscript{14} algorithm applied does not involve a test set, but rather uses a measure referred to as Full Complexity Prediction Squared Error (FCPSE) as the selection criterion. Previous studies have shown this measure to be effective in selecting the best predicting model (Baker & Richards, 2000). Table 2 displays the Training and Test Set aggregate and Production Set descriptive statistics.

2.3. Comparison of prediction accuracy

Just as each algorithm uses some measure of prediction accuracy for model selection, this study uses predic-

\textsuperscript{10}For an extensive discussion of the derivation of GRNN, see Specht (1991).

\textsuperscript{11}Most thoroughly described by Madala and Ivakhnenko (1994) and Farlow (1984).

\textsuperscript{12}NEUROSHELL \textsuperscript{2} (v3.0) distributed by Ward Systems Group (WSG), Frederick, MD ( )

\textsuperscript{13}For a more detailed discussion see Murphy, Fogler and Kohler (1994), who display that the shape of this curve is “U”-shaped when applying standard backpropagation neural net-

\textsuperscript{14}GMDH-type algorithms like that included in NEUROSHELL \textsuperscript{2} formally differ from true GMDH algorithms in that true GMDH neural networks involve test set extraction and inductive model fitting (see Madala & Ivakhnenko, 1994).
Table 1
Data samples and estimation procedures

<table>
<thead>
<tr>
<th>Model type</th>
<th>Training set</th>
<th>Test set</th>
<th>Production set</th>
<th>Estimation procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear regression</td>
<td>n=163</td>
<td></td>
<td>n=20</td>
<td>OLS</td>
</tr>
<tr>
<td>Backpropagationb</td>
<td>n=131</td>
<td>n=32</td>
<td>n=20</td>
<td>Iterative with momentum termc</td>
</tr>
<tr>
<td>Generalized regression</td>
<td>n=131</td>
<td>n=32</td>
<td>n=20</td>
<td>Genetic (gene pool size=300)d</td>
</tr>
<tr>
<td>GMDH</td>
<td>n=163</td>
<td></td>
<td>n=20</td>
<td>FCPSEe</td>
</tr>
</tbody>
</table>

a Initially selected at random, the same production set was extracted for each model.


c The use of a momentum term allows the network to learn more quickly by adjusting weights according to the direction of the error from the previous iteration, rather than simply applying another random set of weights.

d The gene pool for each generation of equations can range from 20 to 300. Increasing the gene pool substantially increases training time, but provides greater probability of identifying a best predicting equation.

e Referred to as Full Complexity Prediction Squared Error, FCPSE is an algorithm for which WSG retains proprietary rights. In general, FCPSE is based on a measure of training set squared error, an overfit penalty (see Ward Systems Group, 1995, p. 149) and an additional penalty for model complexity.

Table 2
Descriptive statistics for data samples

<table>
<thead>
<tr>
<th>Variable</th>
<th>Training and test set (N=163)</th>
<th>Production set (N=20)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>CV (%)</td>
</tr>
<tr>
<td>Average enrolment per grade level</td>
<td>33</td>
<td>88.7</td>
</tr>
<tr>
<td>Length of school day (h)</td>
<td>176.3</td>
<td>1.2</td>
</tr>
<tr>
<td>Length of school year (days)</td>
<td>6.5</td>
<td>9.3</td>
</tr>
<tr>
<td>Average class size</td>
<td>17.7</td>
<td>19.1</td>
</tr>
<tr>
<td>Percentage of student identified for special education</td>
<td>10.1</td>
<td>42.4</td>
</tr>
<tr>
<td>Student-to-computer ratio</td>
<td>8.26</td>
<td>88.6</td>
</tr>
<tr>
<td>Percentage of parents with postsecondary education</td>
<td>22.6</td>
<td>44.6</td>
</tr>
<tr>
<td>Percentage of families in poverty</td>
<td>13.3</td>
<td>63.2</td>
</tr>
<tr>
<td>Wealth index</td>
<td>0.19</td>
<td>283.9</td>
</tr>
<tr>
<td>Instructional expenditures per pupil</td>
<td>3838</td>
<td>18.2</td>
</tr>
<tr>
<td>Percentage of students passing maths competencies (Grade 4)</td>
<td>44.1</td>
<td>25.9</td>
</tr>
</tbody>
</table>

...tion accuracy on the production set to determine which architectures should be further explored for their usefulness in our educational productivity analyses. This does not mean, however, that the potential of other architectures should or will be entirely ignored. Each model including the linear regression model is compared on a measure of Mean Absolute Percentage Error (MAPE) of prediction on the production set schools.

2.4. Drawing inferences from the models

In econometrics we have a variety of standard, generally accepted methods at our fingertips for drawing inference from estimated models. Neural networks and genetic algorithms present several difficulties in this regard, for which there are currently no well-accepted standards. The GMDH method provides the potential advantage of yielding additive polynomial regression equations that may ultimately be explored using our current inferential tool-kit (Liao, 1992). Following GMDH estimation, parameters identified by GMDH are tested for significance using OLS methods and traditional statistical software (SAS v6.2).

The degrees of complexity of weighting matrices, replication of inputs in multiple slabs seemingly producing massive collinearity and the sigmoid rescaling...
(squashing) of data, in backpropagation, make traditional interpretation difficult if not impossible.\textsuperscript{15} Similar conditions exist with respect to the interpretation of GRNN parameters. One response to these difficulties is to perform sensitivity analyses using trained neural networks (deterministic non-linear models), rather than try to dissect individual network parameters. Sensitivity analysis is a useful method for characterizing the response of a dependent variable to changes or differences in levels in individual or multiple input measures. Kaplan and Elliott (1997) apply sensitivity analysis with a hierarchical structural model to determine the aggregate of direct and indirect effects of manipulating policy levers on student-level science achievement. While not commonplace in educational research or social science research in general, neural network sensitivity analysis is used with increasing regularity in medicine (Casciani & Parham, 1998; Parham, 1998; Reid, Nair, Kashani & Rao, 1994) and in engineering and operations research (Sharda & Wang, 1996).

In this study, a comparative sensitivity simulation was run using each of the three algorithms and the linear regression model. This method, like the testing of parameters identified by GMDH, by necessity is deductive in that it involves selecting a variable or variables of interest to be manipulated. For the simulation, class size was chosen as the policy lever to be manipulated. Class size was chosen for its growing prevalence in policy debates (Betts, 1996; Card & Krueger, 1996). Class sizes were incrementally (by 2) adjusted for production set \( n=20 \) schools from a minimum of 10 to a maximum of 20. With the class sizes adjusted, production set data were re-entered into trained networks and estimated regression equations. Production set mean productivity levels, given adjusted class size, were then determined.

2.5. Potential shortcomings of the oversimplified application

One significant shortcoming of the combination of methods and data used in this study is that they fail to significantly advance our understanding of causality in educational productivity. The current body of literature on educational productivity suggests the relative importance, if not absolute necessity, of (1) the use of lagged measures of inputs and outputs in order to estimate the change in \( Y \) with respect to the change in \( X \) and (2) accounting for the presence of endogenous variables by using instrumental variables (IV) methods such as two-stage least-squares (2SLS). Regarding inclusion of lagged measures, the Vermont School Reports had not been in existence for enough years at the time of this study for the fourth graders initially measured to have reached their second point of measurement in the eighth grade. In addition, differences in cohort performance across schools rendered pseudo value added\textsuperscript{16} methods an unreasonable proxy.\textsuperscript{17}

Currently, the question of how to deal with endogenous variables in the context of neural network estimation is somewhat more problematic. It seems reasonable that indirect or two-stage least-squares methods might be simulated by generating sets of predicted values for endogenous variables using neural networks trained on sets of predetermined instruments and that those predicted values could be used in subsequent training of second-stage neural networks. An alternative would be to apply the new insights on potential functional forms generated by single-stage neural networks as parameters to be tested in the context of a more conventional simultaneous equation model. This study stops short of undergoing either of these endeavors and focuses solely on non-linear relationships that may emerge in the single-stage, fully contemporaneous, cross-sectional model. Attempts are ongoing to validate the statistical properties of the outcomes of the two-stage neural network approach and the relative usefulness of each of the two alternatives presented.

3. Results

3.1. Comparisons of prediction accuracy

Table 3 displays the training set and production set fit (R-squared) and production MAPE values. Note that the linear regression models provide a particularly weak fit to the training set, but actually perform quite well on both fit to and prediction of the production set. GRNN, on the other hand, appears to have significantly overfit the training set, despite the cross-validating estimation procedure, causing a failure to predict accurately or fit the production set. Backpropagation provides a moderate fit to both the training and production sets, but does not perform as well as the linear model on prediction accuracy. GMDH yields the most consistent fit to both training and production sets and produces accurate predictions.

\textsuperscript{15} For a discussion of how these factors in particular confound our traditional econometric understanding of backpropagation, see McMenamin (1997).

\textsuperscript{16} For example, differencing current-year eighth grade student performance with current-year fourth grade student performance, as done in a number of other studies including Ferguson and Ladd (1996).

\textsuperscript{17} It is recognized that this cohort-to-cohort variance across schools will also result in changes to estimated models from year to year. But, while problematic, it is again emphasized that the objective of this study is not to derive a conclusive model of schooling productivity in Vermont, but to explore the potential relationships revealed by neural networks.
Table 3
Measures of fitness and prediction accuracy

<table>
<thead>
<tr>
<th></th>
<th>Training $R^2$ (n=163)</th>
<th>Production $R^2$ (n=20)</th>
<th>MAPE$^a$</th>
<th>SD of MAPE$^b$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Untreated models</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear</td>
<td>0.09</td>
<td>0.26</td>
<td>18.0</td>
<td>0.154</td>
</tr>
<tr>
<td>Backpropagation</td>
<td>0.15</td>
<td>0.23</td>
<td>20.5</td>
<td>0.187</td>
</tr>
<tr>
<td>GRNN</td>
<td>0.71</td>
<td>0.10</td>
<td>18.6</td>
<td>0.165</td>
</tr>
<tr>
<td>GMDH</td>
<td>0.38</td>
<td>0.53</td>
<td>13.8</td>
<td>0.126</td>
</tr>
<tr>
<td><strong>Log (ln) transformed</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear</td>
<td>0.09</td>
<td>0.39</td>
<td>14.9</td>
<td>0.140</td>
</tr>
<tr>
<td>Backpropagation</td>
<td>0.12</td>
<td>0.18</td>
<td>17.4</td>
<td>0.154</td>
</tr>
<tr>
<td>GRNN</td>
<td>0.76</td>
<td>0.09</td>
<td>24.3</td>
<td>0.201</td>
</tr>
<tr>
<td>GMDH</td>
<td>0.44</td>
<td>0.36</td>
<td>14.4</td>
<td>0.115</td>
</tr>
</tbody>
</table>

$^a$ MAPE=Mean absolute percentage error.
$^b$ SD=Standard deviation of the mean of the absolute values of the error (percentage error) terms.

Table 4
Linear regression model estimates (* indicates $p<0.05$)

<table>
<thead>
<tr>
<th></th>
<th>Untreated data</th>
<th>Log transformed</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>SE</td>
<td>VIF</td>
<td>Estimate</td>
<td>SE</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.62</td>
<td>0.83</td>
<td>–</td>
<td>−0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>Average enrolment per grade level</td>
<td>0.01</td>
<td>0.02</td>
<td>1.24</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>Length of school year (days)</td>
<td>−0.04</td>
<td>0.81</td>
<td>1.12</td>
<td>−0.02</td>
<td>0.72</td>
</tr>
<tr>
<td>Length of school day (h)</td>
<td>0.42</td>
<td>0.23</td>
<td>1.13</td>
<td>0.70</td>
<td>0.41</td>
</tr>
<tr>
<td>Average class size</td>
<td>−0.03</td>
<td>0.11</td>
<td>1.27</td>
<td>−0.06</td>
<td>0.10</td>
</tr>
<tr>
<td>Percentage of students identified for special education</td>
<td>−0.02</td>
<td>0.05</td>
<td>1.19</td>
<td>−0.01</td>
<td>0.05</td>
</tr>
<tr>
<td>Student-to-computer ratio</td>
<td>−0.02</td>
<td>0.02</td>
<td>1.14</td>
<td>−0.06</td>
<td>0.04</td>
</tr>
<tr>
<td>Percentage of parents with postsecondary education</td>
<td>0.12*</td>
<td>0.05</td>
<td>1.51</td>
<td>0.10*</td>
<td>0.05</td>
</tr>
<tr>
<td>Percentage of families in poverty</td>
<td>−0.03</td>
<td>0.03</td>
<td>1.58</td>
<td>−0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>Wealth index</td>
<td>−0.02</td>
<td>0.04</td>
<td>1.40</td>
<td>−0.02</td>
<td>0.07</td>
</tr>
<tr>
<td>Instructional expenditures per pupil</td>
<td>0.03</td>
<td>0.11</td>
<td>1.22</td>
<td>0.02</td>
<td>0.11</td>
</tr>
</tbody>
</table>

$R$-squared=0.10
Adjusted $R$-squared=0.05

3.2. Structure of the “winning” GMDH models

Given that the output of the GMDH algorithm is a complex additive polynomial, with two- and three-way interaction terms, it is perhaps worth exploring the significance of these terms with respect to that of the terms in the linear model. Table 4 displays the estimates for the linear model including all 10 variables, with variance inflation factors (VIFs) reported to indicate degrees of multicollinearity.$^{18}$ These estimates are derived by refitting the equations to the full, rather than subdivided sample ($n=183$). In the full linear model, only the parent level of education variable displays significance. The total variance explained in the outcome measure remains relatively low (10%) with the adjusted $R$-squared at only 5%.

Tables 5 and 6 display the significant coefficients from the higher-order and interaction terms identified by GMDH. Some non-significant lower-order terms and terms underlying two- or three-way interactions are also reported. Full GMDH equations are displayed in Appendix B. Coefficients in Tables 5 and 6 are the result of re-estimation of the GMDH equation terms using the full data set ($n=183$) and standard statistical software (sas version 6.2). While GMDH involves a complex linear re-scaling of data prior to estimation,$^{19}$ data for this analysis

$^{18}$ Where the standard threshold for multicollinearity is VIF>5 (or tolerance<0.2).

$^{19}X'=(X−\text{Min})/[(\text{Max}−\text{Min})−1]$, further described in Ward Systems Group (1995), p. 158.
Table 5
Significant \((p<0.05)\) estimates from non-linear model “A”. \(R\)-squared=0.45, adjusted \(R\)-squared=0.28 (full model) \((n=183)\)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>(t)-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1444</td>
<td>-3.2</td>
</tr>
<tr>
<td>Enrolment</td>
<td>0.40</td>
<td>2.8</td>
</tr>
<tr>
<td>Length of school year</td>
<td>4540</td>
<td>3.2</td>
</tr>
<tr>
<td>((\text{Length of school year})^2)</td>
<td>-4872</td>
<td>-3.3</td>
</tr>
<tr>
<td>((\text{Length of school year})^3)</td>
<td>1778</td>
<td>3.4</td>
</tr>
<tr>
<td>Length of school day</td>
<td>3.06</td>
<td>2.3</td>
</tr>
<tr>
<td>Average class size</td>
<td>258</td>
<td>2.0</td>
</tr>
<tr>
<td>((\text{Average class size})^2)</td>
<td>-87.1</td>
<td>-1.4</td>
</tr>
<tr>
<td>((\text{Average class size})^3)</td>
<td>-3.41</td>
<td>-2.1</td>
</tr>
<tr>
<td>Student to computer ratio</td>
<td>-0.46</td>
<td>-2.9</td>
</tr>
<tr>
<td>Percentage of families in poverty</td>
<td>-15.5</td>
<td>-4.0</td>
</tr>
<tr>
<td>Wealth index</td>
<td>2.13</td>
<td>2.1</td>
</tr>
</tbody>
</table>

Two-way interaction terms

<table>
<thead>
<tr>
<th>Term</th>
<th>Estimate</th>
<th>(t)-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Special education (\times) parent education</td>
<td>21.5</td>
<td>2.1</td>
</tr>
<tr>
<td>School year (\times) poverty</td>
<td>15.2</td>
<td>3.9</td>
</tr>
<tr>
<td>School year (\times) class size</td>
<td>-268</td>
<td>-2.1</td>
</tr>
<tr>
<td>School day (\times) wealth</td>
<td>-2.11</td>
<td>-2.1</td>
</tr>
<tr>
<td>Enrolment (\times) special education</td>
<td>-0.16</td>
<td>-2.8</td>
</tr>
</tbody>
</table>

Three-way interaction terms

<table>
<thead>
<tr>
<th>Term</th>
<th>Estimate</th>
<th>(t)-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>School year (\times) special education (\times) parent education</td>
<td>-21.6</td>
<td>-2.1</td>
</tr>
</tbody>
</table>

\(a\) For full model, including all parameter estimates, see Appendix B.

\(b\) Non-significant term included as underlying term to significant cubed term.

\(c\) Consists of terms not significant as individual predictors.

---

Table 6
Significant \((p<0.05)\) estimates from non-linear model “B”. \(R\)-squared=0.44, adjusted \(R\)-squared=0.27 \((n=183)\)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>(t)-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.06</td>
<td>-1.46</td>
</tr>
<tr>
<td>Enrolment per grade level</td>
<td>-0.02</td>
<td>-0.53</td>
</tr>
<tr>
<td>((\text{Enrolment per grade level})^2)</td>
<td>-0.03</td>
<td>-1.28</td>
</tr>
<tr>
<td>((\text{Enrolment per grade level})^3)</td>
<td>0.04</td>
<td>2.28</td>
</tr>
<tr>
<td>Length of school day</td>
<td>0.99</td>
<td>2.00</td>
</tr>
<tr>
<td>Percentage of families in poverty</td>
<td>-0.17</td>
<td>-3.55</td>
</tr>
<tr>
<td>Wealth index</td>
<td>-0.12</td>
<td>-1.10</td>
</tr>
<tr>
<td>((\text{Wealth index})^2)</td>
<td>0.70</td>
<td>2.48</td>
</tr>
<tr>
<td>((\text{Wealth index})^3)</td>
<td>-0.41</td>
<td>-2.23</td>
</tr>
<tr>
<td>Length of school year</td>
<td>-5.86</td>
<td>-1.28</td>
</tr>
<tr>
<td>Special education</td>
<td>-0.06</td>
<td>-0.76</td>
</tr>
</tbody>
</table>

Two-way interaction terms

<table>
<thead>
<tr>
<th>Term</th>
<th>Estimate</th>
<th>(t)-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length of school year (\times) poverty</td>
<td>11.5</td>
<td>3.31</td>
</tr>
<tr>
<td>Length of school year (\times) parent education</td>
<td>5.85</td>
<td>1.29</td>
</tr>
<tr>
<td>Parent education (\times) poverty</td>
<td>-0.03</td>
<td>-0.44</td>
</tr>
<tr>
<td>Enrolment per grade level (\times) special education</td>
<td>-0.21</td>
<td>-3.16</td>
</tr>
<tr>
<td>Enrolment per grade level (\times) length of school year</td>
<td>-5.40</td>
<td>-2.20</td>
</tr>
</tbody>
</table>

Three-way interaction terms

<table>
<thead>
<tr>
<th>Term</th>
<th>Estimate</th>
<th>(t)-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length of school year (\times) parent education (\times) poverty</td>
<td>-19.3</td>
<td>-2.59</td>
</tr>
</tbody>
</table>

\(a\) Included as underlying term to significant higher-order terms.

\(b\) Included as underlying terms to significant two-way interaction.

\(c\) Included as underlying two-way interactions to significant three-way interaction.
were linearly re-scaled around as a percentage of their median, around a median of “1”. For starters, the non-linear models explain far more of the total variance (R-squared=0.44 to 0.45) than their linear counterparts and despite the high number of parameters display reasonably high adjusted R-squared values (0.27 to 0.28).

GMDH polynomials produced numerous statistically significant relationships including a third-order polynomial relationship for the length of school year variable in the non-logged model and a third-order polynomial relationship for the school size variable (average enrollment per grade level) in the logged model. Some similar two-way interactions appear in both models, including the interaction of length of school year and poverty and the interaction of school size and special education proportions, for which there may be theoretically reasonable explanations. For example, it seems reasonable to assume that an extended school year, by its custodial and peer group effects, may provide a more stimulating and supportive environment for learning for economically deprived students. On the other hand, more affluent students might actually seek even greater academic challenges than their local public schools provide once the school doors close for the summer.20 Given the limited sample and school aggregate data under investigation, any such interpretation is highly speculative. Yet, that such questions were raised by GMDH suggests the need for further exploration of these interactions with more detailed and more comprehensive student-level data.

These significant coefficients, however, should be considered “questions raised” rather than answers given. From formal statistical pedagogy, for example, we would be required to address the significance of the parameters in the context of all of the possible hypotheses that we were testing. In this case those hypotheses include all possible second- and third-order terms of the 10 original inputs, and all possible two- and three-way interactions, although not all were included in the resultant GMDH models.

3.3. Inference by sensitivity analysis

The results of the simulation are displayed in Fig. 1. Note that, for the range of class sizes from 10 to 20, the linear model produces a gradually declining constant slope between the ranges of 44 and 46% passing maths competencies. The flexible non-linear models produce responses ranging from a very slight curvature (backpropagation) to similarly dramatic curvatures (GMDH and GRNN), suggesting substantial performance gains for class sizes declining from 14 to 10. Both GMDH and GRNN as well indicate performance improvement from a class size of 18 through 20, while the linear model continues its gradual downward pattern. The consistency of these results warrants further exploitation of this relationship. That a pattern of such consistency emerged from the data, given the variety of possibilities for non-linear model fitting, provides some evidence of the usefulness of these methods.

4. Conclusions and implications

The computer-assisted inductive methods used in this experiment yielded differences in functional form and provided new direction for parameter selection when compared with typical linear regression methods. The non-linearities and interactions identified, although theoretically reasonable in many cases, might not necessarily be expected based on the current literature on educational productivity; thus, it is unlikely that we would find such patterns by traditional deductive methods. The GMDH method in particular deserves further attention, as it consistently produces models with better overall fitness and greater prediction accuracy, as well as yielding more easily interpreted outcomes. The emphasis on the value of GMDH, however, does not discount the usefulness of other architectures which, for example, in the sensitivity simulation provided validation of the curvilinear relationship revealed by GMDH.

As an extension of the earlier work of Figlio (1999), this study validates not only our need to test less restrictive, more flexible functional forms for understanding effect magnitudes in educational production, but the usefulness of inductive algorithms for revealing potential relationships to be further explored with these functional forms. Interestingly, the GMDH method of polynomial expansion applied to data in logged form may be particularly useful for identifying parameters to be tested via the transllog specification discussed by Figlio.21

The intent of this study is by no means to promote radical inductionism as a sole means to better understanding educational productivity. This is a frequent

---

20 Baker and Richards (1998) document the flight of gifted and talented pupils to a handful of rapidly growing, but relatively costly, private summer academic programs from 1984 through 1994.

21 Similar to GMDH polynomial estimation applied to logged data, transllog involves a geometric expansion of the logarithm, providing a local second-order approximation to any production function (Figlio, 1999, p. 243). Figlio’s transllog education production function takes the form: \( \ln y_t = \alpha_0 + \sum_j \beta_j (\ln X_{it}) + 1/2 \sum_k \gamma_k (\ln X_{it})^2 + \sum \delta_j (\ln X_{it-1}) (\ln X_{it}) + \sum \theta_j (\ln X_{it-1}) (\ln y_{it-1}) + \alpha_0 (\ln y_{it-1}) + 1/2 \sigma_0 (\ln y_{it-1})^2 \) for all \( j \neq k \), where \( X_{it} \) includes the school input variables, the peer variables and the family variables in the model at time \( t-1 \). Note that Figlio’s model, unlike the models generated herein, estimates value added by inclusion of \( y_{it-1} \), the lagged measure of the outcome, on the right-hand side of the equation.
Fig. 1. Sensitivity simulation results — effect of changing class size on maths performance.

criticism of the type of data mining methods applied herein, which is warranted to the extent to which these methods, while often raising interesting questions, may also identify a variety of spurious relationships that advance prediction accuracy despite questionable meaning. Rather, the objective of this exercise is to promote greater balance in our use of deductive and inductive tools, and to present a framework for using flexible non-linear analysis as an inductive complement to traditional production function analysis. The value of any of the inductive, exploratory methods presented herein only becomes apparent when complemented with deductive methods. We contend that the converse of this statement is also true.

Acknowledgements

The author wishes to thank the many reviewers who have formally provided feedback on earlier versions of this article and various colleagues and friends of colleagues, including Craig Richards and Jane Monroe of Teacher College, Columbia University, Suzanne Rice of The University of Kansas, and Bruce Cooper of Fordham University who informally provided feedback on earlier versions of this article and related works.

Appendix A. Variable descriptions

Average Enrolment per Grade Level (ENGRD) — derived using variable ENROLL98 (Vermont State Reports 1998) divided by number of grade levels in the school.

Length of School Day (h) — variable LSD98 (Vermont State Reports 1998).

Length of School Year (days) — variable LSY98 (Vermont State Reports 1998).

Average Class Size — variable AVGCS98 (Vermont State Reports 1998).

Proportion of Students Receiving Special Education Services — variable SPECED97 (1997 reported data) (Vermont State Reports 1998).


Proportion of Parents with Postsecondary Education (HIGHED) — determined as the sum of EDUC690 (1990 census data, Vermont State Reports 1998), the per-
Appendix B. Full GMDH equations

B.1. Model A — linear GMDH scaling only

Best formula:
\[ Y = -0.17 - 0.7X2 - 4.8 \times 10^{-2}X1 + 0.15X3 - 0.32X6 + 9.3 \]
\[ \times 10^{-2}X7 - 0.25X8 - 7.6 \times 10^{-2}X4 - 5.4 \times 10^{-2}X5 + 0.21X4^2 \]
\[ - 0.18X5^2 - 0.35X4X5 + 0.3X4X7 - 0.44X5X7 + 2X2^2 \]
\[ - 0.13X8^2 + 0.33X2^3 + 0.24X8^3 + 1.5X2X8 - 0.19X2X4 \]
\[ - 0.13X2X5 + 0.58X2X7 + 0.32X2X4^2 - 0.46X2X5^2 \]
\[ - 0.68X2X4X5 + 0.55X2X4X7 - 1.1X2X5X7 + 0.32X6^2 \]
\[ + 0.12X7^2 - 6.1 \times 10^{-2}X6^2 + 0.27X6X7 + 0.2X4X10 \]
\[ - 0.49X3X9 - 0.4X1X5 + 0.18X5X6 - 0.24X2X10 \]
\[ + 0.15X4X8 + 5.9 \times 10^{-2}X1^2 - 7.4 \times 10^{-2}X4^3 - 0.25X1X4. \]

Legend:
\[ X1 = 2(\text{ENGRD} + 25.84)/117.12 - 1 \]
\[ X2 = 2(\text{LSY98} - 167.41)/17.15 - 1 \]
\[ X3 = 2(\text{LSD98} - 5.37)/2.29 - 1 \]
\[ X4 = 2(\text{AVGCS98} - 10.68)/13.97 - 1 \]
\[ X5 = 2(\text{SPECED97} - 1.74)/16.65 - 1 \]
\[ X6 = 2(\text{STUCMP97} + 5.72)/27.87 - 1 \]
\[ X7 = 2(\text{HIGHED} - 2.51)/40.27 - 1 \]
\[ X8 = 2(\text{PPOV96} + 3.23)/32.87 - 1 \]
\[ X9 = 2(\text{WEALTH} + 0.88)/2.14 - 1 \]
\[ X10 = 2(\text{INSPP} - 2354.86)/3001.95 - 1 \]
\[ Y = 2(\text{PASS_MATH} - 21.47)/44.9 - 1 \]

B.2. Model B — log transformation and GMDH scaling

Best formula:
\[ Y = -0.92X2 - 3.6 \times 10^{-2}X1 + 0.19X7 - 0.18 + 0.18X3 \]
\[ - 0.25X8 - 7.4 \times 10^{-2}X4 - 7.7 \times 10^{-2}X5 + 1 \times 10^{-2}X6 + 8 \]
\[ \times 10^{-2}X10 + 2.2X2^2 + 0.32X2^3 + 0.37X8^3 + 0.64X2X7 \]
\[ + 1.6X2X8 - 0.29X7X8 - 2.6X2X7X8 - 0.11X1^2 - 0.17X5^2 \]
\[ + 0.15X6^2 + 0.18X1^3 - 0.28X5^3 - 0.82X1X5 + 0.33X5X6 \]
\[ + 0.62X1X5X6 + 0.24X^2 - 0.1X4^3 - 7.2 \times 10^{-2}X6^3 \]
\[ + 0.29X6X10 + 0.29X9^2 - 0.14X9^3 - 0.73X3X9 - 1.9X2X3 \]
\[ + 0.24X4X7 - 0.64X2X4 - 0.42X2X10 + 0.23X8^2 \]
\[ - 0.4X1X2 + 0.15X6X9 + 0.11X4X9 + 0.12X4X8 + 7.7 \]
\[ \times 10^{-2}X5X10. \]

Legend:
\[ X1 = 2(\text{ENGRD} - 1.58)/3.19 - 1 \]
\[ X2 = 2(\text{LSY98} - 5.11)/0.11 - 1 \]
\[ X3 = 2(\text{LSD98} - 1.58)/0.59 - 1 \]
\[ X4 = 2(\text{AVGCS98} - 2.44)/0.83 - 1 \]
\[ X5 = 2(\text{SPECED97} - 1.28)/1.87 - 1 \]
\[ X6 = 2(\text{STUCMP97} - 0.88)/2.12 - 1 \]
\[ X7 = 2(\text{HIGHED} - 2.01)/2.0 - 1 \]
\[ X8 = 2(\text{PPOV96} - 0.78)/3.11 - 1 \]
\[ X9 = 2(\text{WEALTH} + 0.58)/1.36 - 1 \]

22 Presented as generated by neuroshell 2. Equations (trained networks) can either be generated as above for browsing (but not application due to reduction of precision as a result of rounding) or generated as run-time applications in visual basic or C++ to be called by other programs including custom applications in MS Excel. Such use is common in sensitivity simulation.
\[ X_{10} = 2(\text{INSPP} - 7.86)/0.75 - 1 \]
\[ Y' = 2(\text{PASS}_\text{MATH} - 3.24)/1.02 - 1 \]

References


