# Educational Performance Indicators and LEA League Tables

GEOFFREY WOODHOUSE & HARVEY GOLDSTEIN

ABSTRACT *A common procedure for using examination results as performance indicators is based upon residuals from regression analysis. These analyses are typically applied to aggregated data: this paper demonstrates that such procedures give unstable results. It is suggested that aggregate-level analyses are uninformative and that useful comparisons cannot be obtained without employing multilevel analyses using student-level data.*

## INTRODUCTION

In recent years there has been a considerable interest in devising indicators of school 'performance' and using these in analyses to compare the 'efficiency' of individual schools or local authorities. An early attempt by the Inner London Education Authority (ILEA, 1980) compared the average public examination results in one year of the authority's secondary schools. Subsequently, the Department of Education and Science (DES, 1983, 1984), published comparisons of average local education authority (LEA) examination results, and the same data have been re-analysed, for example, by Gray & Jesson (1987) and by Levitt & Joyce (1987).

All these analyses have dealt with comparisons between 'aggregate level' units, whether schools or LEAs, after adjusting for pre-existing differences, and using data at the same aggregate level. For example, the ILEA carried out a regression analysis using the average (16-year) school examination result as the outcome or 'response variable'. The 'input' measures made at 11 years included for each school the proportion of pupils in two bands of verbal reasoning scores (VRQ) and average socio-economic status. A 'residual deviation score' was then assigned to each school, being the difference between the school's actual examination score and that predicted by the regression equation. Schools were then ranked on this difference, interpreted as a measure of school 'efficiency', i.e. the difference between actual and 'expected' performance. The analyses using LEA aggregated data have followed similar procedures, typically using various measures of educational expenditure and social background to adjust for pre-existing differences.

The purpose of this paper is to show that these procedures have little justification in theory or in practice: that there are severe difficulties in attaching a causal interpretation to the residuals as measures of efficiency and that in practice the rankings show considerable instability when the model used to produce them is subjected to minor changes.

A variant of these procedures is that used by Jesson *et al.* (1987) based on a technique known as Data Envelopment Analysis (DEA). This also uses aggregated data and considers a ratio of outcome to input variables rather than a regression residual as a measure of efficiency. This procedure also is open to serious objections which we discuss in an appendix.

## UNITS OF ANALYSIS

All the above analyses choose an aggregate level unit, the school or the local authority, as the basic unit in the analysis. The use of these units, however, does not allow us to study within-unit relationships. For example, school A may perform better than school B for students with low intake test scores but worse than school B for students with high intake test scores. If we know the within-school relationship for individual students between examination scores and their input scores, we can in principle compare each school's outcome score with its predicted score for students with specific values of the input score. If, however, aggregate data only are available, detailed study based on individual student characteristics is impossible to carry out.

Analyses which use only LEA or school mean data thus have a rather limited interpretation. In particular, it is difficult to see how a causal interpretation can be based upon a study of the relationship between mean outcomes and mean inputs. Within any school, students will have a range of values on the input variables as well as the outcome variables. If the relationship for students (of outcome to input variables) varies from school to school, any analysis based on school means cannot provide information on this. Furthermore, even if the within-school relationship is constant, it may still be quite different from the relationship between-schools using the same variables aggregated to the school level. For a discussion of these issues see Aitkin & Longford (1986) and Goldstein (1987, chapters 2 and 3).

There are further serious problems with aggregate-level analyses which arise from the typical sensitivity of such analyses to the mathematical and statistical assumptions built into them.

First, different assumptions produce different choices of input variables (or 'predictors') for inclusion in the analysis. Thus for example the DES (1984), using step-wise multiple regression on aggregate LEA data, produced models involving nine or more predictors. Gray & Jesson (1987) in their re-analysis of the same data used only four predictors. The four were not a subset of the nine used by the DES to predict the same outcome variable, and a fifth variable, not considered by the DES, was later added to adjust for independent schooling. Levitt & Joyce (1987) used principal component analysis to produce a further, different, set of predictors from the same data set. These models were comparable in the degree to which they fitted the data according to the usual statistical criteria. Yet using their residuals for ranking produced a markedly different position for some LEAs.

Secondly, the rank order of the residuals depends on whether or not they are 'standardised'. Residuals arising from the analysis will be differently distributed, in particular with different variances, at different points in the data space, and there is a case for dividing each one by its estimated standard deviation before making any comparison. Analyses have differed in their procedures. Thus the DES (1983, 1984) published tables of actual and predicted scores without the information needed to standardise the residuals. Gray & Jesson produced league positions based on the DES analyses using 'raw' residuals (1987, models 6 and 7) without commenting on the possible effects of standardisation, although they standardised the residuals from their own analysis (model 8). Levitt & Joyce did not standardise.

Thirdly, the rank order of the residuals, even if 'standardised', depends not only on the choice of variables but also on the precise form in which these appear in the model. In the analyses we have cited neither the DES nor Gray and Jesson appear to have considered non-linear transformations of their input variables, while Levitt & Joyce appear to have considered only a logarithmic transformation of response and explanatory variables. The fact, however, that the examination results and the socio-economic variables used in these models are measured on arbitrary scales suggests that other transformations could be considered.

In the remainder of this paper we give the results of a re-analysis of one particular model, that of Gray & Jesson (1987, model 8). Allowing that the procedure they used in order to arrive at their five input variables was an appropriate one within the limitations of step-wise regression, we show how the use of non-linear transformations of these variables and of the response variable improves the *statistical fit* in different ways depending on the combination used. None of the resulting models, however, can be shown to be 'the best' and each gives rise to a different rank order of the residuals. We argue that any similar model will show instability of residual ranking when subjected to similar routine manipulation, and conclude that residuals from aggregate-level regression analyses are inherently unreliable as measures of efficiency.

## RE-ANALYSIS PRELIMINARIES

The model produced by Gray & Jesson may be expressed as follows:

$$y_i=\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+\beta_3x_{3i}+\beta_4x_{4i}+e_i \qquad (1)$$

where for the ith LEA (i=1, 2, . . ., 96),

$y_i=z_{1i}+0.5z_{2i}$;

$z_{1i}$= percentage of maintained school leavers achieving at least five O-level passes at grade A, B or C or CSE passes at grade 1;

$z_{2i}$= percentage of children attending independent schools in the LEA as day pupils;

$x_{1i}$= percentage of children living in households whose head is a non-manual worker, excluding junior non-manual workers and non-manual supervisors (a 'high socio-economic group' indicator);

$x_{2i}$= percentage of children living in households whose head is a semi-skilled or unskilled manual worker, personal service or farm worker (a 'low socio-economic group' indicator);

$x_{3i}$= percentage of children living in one-parent families;

$x_{4i}$= percentage of children born outside the UK, Ireland, USA and the Old Commonwealth or in households whose head was born outside the UK, Ireland, USA and the Old Commonwealth;

$e_i$= residual term. In the standard linear model this is assumed to have a zero mean and constant variance $\sigma^2$;

$\beta_0, \ldots, \beta_4$ are population parameters to be estimated by fitting the model to the data. In the standard linear model these data are assumed to refer to a random sample from a larger population, which is not the case here. We return to this point later.

It will be seen that the response variable $y$ is the percentage of leavers with five or more O-level equivalents adjusted for independent schooling in each LEA by adding half the percentage of children attending independent schools in the LEA as day pupils. The coefficient of $z_2$ was fixed at 0.5 by Gray & Jesson on the grounds that at least half those attending independent schools could be expected to gain five or more O-level equivalents.

The four indicators $x_1$, $x_2$, $x_3$ and $x_4$ were found according to Gray & Jesson 'using an approach based on regression analysis' (p. 36). They provide no further details of their method and we do not explore it. Rather, we shall be concerned to examine:

(a) the 'goodness of fit' of Gray & Jesson's model to the data; and

(b) the stability of the residual rankings, or 'league table', when the model is subjected to minor perturbations.

Fitting model 1 to the data and ranking the LEAs on the resulting standardised residuals produced a league table very similar to Gray & Jesson's. The differences are given in Table I: Gloucestershire, Ealing and Durham were one place lower on our table than on Gray & Jesson's; Walsall was two places lower. These differences most probably reflect rounding procedures: we are confident that our data set and procedures were essentially equivalent to those used by Gray & Jesson.

TABLE I. *Differences between Gray & Jesson's (1987) model 8 and our model 1*

| LEA | Model 8 position | Model 1 position | Model 8 raw residual (1 d.p.) | Model 1 raw residual (3 s.f.) | Model 1 std residual (3 s.f.) |
|---|---|---|---|---|---|
| Gloucestershire | 24 | 25 | 1.5 | 1.46 | 0.548 |
| Sefton | 25 | 24 | 1.5 | 1.47 | 0.553 |
| Ealing | 33 | 34 | 0.8 | 0.872 | 0.357 |
| Buckinghamshire | 34 | 33 | 1.0 | 0.955 | 0.364 |
| Walsall | 40 | 42 | 0.3 | 0.390 | 0.149 |
| Northumberland | 41 | 40 | 0.4 | 0.409 | 0.155 |
| Rotherham | 42 | 41 | 0.4 | 0.391 | 0.150 |
| Durham | 49 | 50 | −0.2 | −0.158 | −0.0596 |
| Leeds | 50 | 49 | −0.2 | −0.155 | −0.0584 |

Estimates for the parameters of model 1 are given in Table II. Note that throughout we use Greek letters ($\beta_0$, $\beta_1$, $\sigma^2$, etc.) to denote population parameters and corresponding Roman letters ($b_0$, $b_1$, $s^2$, etc.) to denote their estimates from ordinary least-squares regression.

All the model parameters except for $\beta_1$ are poorly estimated, with standard errors in excess of their absolute values.

The estimated correlation matrix of the $b$'s is:

$$\begin{array}{cccccc}
b_0 & 1 & & & & \\
b_1 & -0.95 & 1 & & & \\
b_2 & -0.87 & 0.87 & 1 & & \\
b_3 & -0.18 & -0.02 & -0.29 & 1 & \\
b_4 & -0.12 & -0.05 & 0.08 & -0.55 & 1 \\
 & b_0 & b_1 & b_2 & b_3 & b_4
\end{array}$$

showing high correlation between $b_0$ and $b_1$, $b_0$ and $b_2$, and $b_1$ and $b_2$.

TABLE II. *Parameter estimates and* $R^2$ *for models 1, 2, 3 and 4 (Standard errors in brackets)*

| | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|
| $b_0$ | 3.36 (4.5) | 0.10 (4.5) | 13.7 (6.7) | 5.93 (7.4) |
| $b_1$ | 0.71 (0.08) | 0.69 (0.07) | 0.87 (0.27) | 0.83 (0.27) |
| $b_2$ | 0.07 (0.14) | 0.01 (0.14) | −1.08 (0.64) | −0.65 (0.65) |
| $b_3$ | 0.05 (0.14) | 0.40 (0.18) | −0.02 (0.14) | 0.31 (0.20) |
| $b_4$ | −0.02 (0.03) | 0.29 (0.11) | −0.02 (0.03) | 0.24 (0.12) |
| $b_{34}$ | | −0.02 (0.007) | | −0.02 (0.008) |
| $b_{11}$ | | | −0.004 (0.005) | −0.003 (0.005) |
| $b_{22}$ | | | 0.03 (0.02) | 0.02 (0.02) |
| $s^2$ | 7.2 | 6.7 | 7.0 | 6.7 |
| $R^2$ | 0.83 | 0.84 | 0.83 | 0.84 |

$b_2$ and $b_3$ are positive (although small), contrary to expectation. This is probably the result of using highly intercorrelated predictors. The correlation matrix for the predictors is:

$$\begin{array}{ccccc}
x_1 & 1 & & & \\
x_2 & -0.91 & 1 & & \\
x_3 & -0.44 & 0.52 & 1 & \\
x_4 & -0.05 & 0.08 & 0.52 & 1 \\
 & & & & \\
 & x_1 & x_2 & x_3 & x_4
\end{array}$$

These results suggest that minor changes to the model will result in large changes to individual residuals.

We now examine plots of the residuals.

The overall frequency plot (Fig. 1) does not appear abnormal. Harrow is an outlier, while Barnet, Bromley and Oldham are the only other LEAs with residuals greater in absolute value than 1.96 standard deviations (s.d.).

The plot against predicted scores (Fig. 2), however, shows that most of the LEAs with residuals below about −1 s.d. have predicted scores in the range 22 to 32 (the exceptions are Bromley and Oldham) while most LEAs with residuals above about +1 s.d. have expected scores outside this central range (the exceptions are Wirral and Newcastle). Closer inspection of the residuals in the range −1 to +1 s.d. shows a similar though less pronounced trend. These trends are a further indication of model

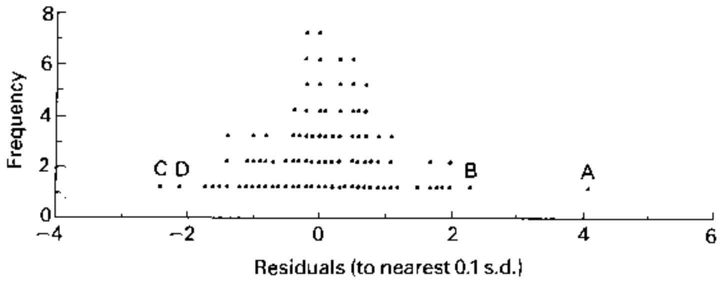

FIG. 1. Frequency plot of model 1 residuals (A = Harrow; B = Barnet; C = Bromley; D = Oldham).

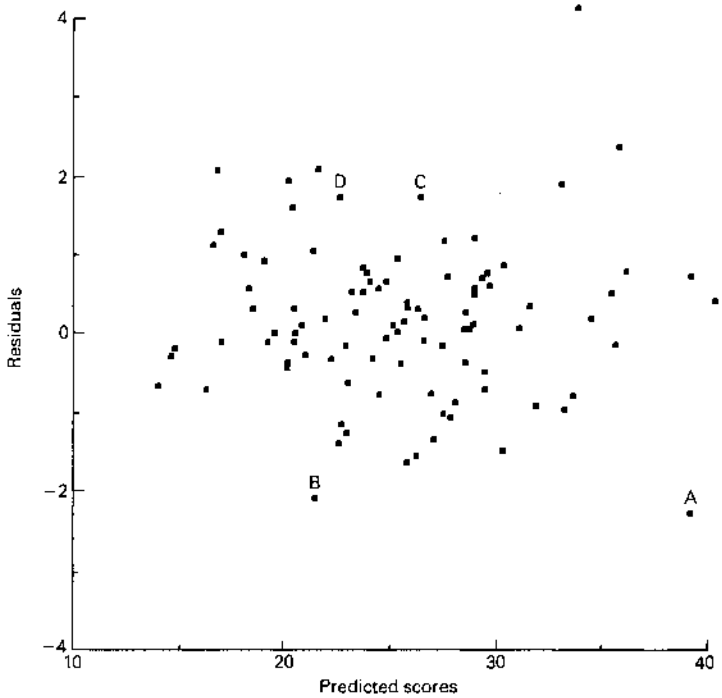

FIG. 2. Model 1 residuals plotted against predicted scores (A = Bromley; B = Oldham; C = Wirral; D = Newcastle).

inadequacy: in particular a need for extra terms in the model (perhaps quadratic or cross-product terms) or for a transformation of the response variable $y$.

The plots against $x_1$ and $x_2$ showed a non-linear trend, less pronounced than in Fig. 2, but nevertheless suggesting that additional quadratic terms in these variables might yield a better fit.

Interpretation of the plots against $x_3$ (Fig. 3) and $x_4$ (Fig. 4) is made difficult by the skewness of the distribution of these variables. It is noticeable, however, that four of the five LEAs with the highest values of $x_3$ have residuals greater in absolute value than 1 s.d. Of these the two with large negative residuals, Brent and Haringey, have high values of $x_4$ while the two with large positive residuals, Liverpool and Manchester, have low values of $x_4$. This suggests that a cross-product term might usefully be included to improve the fit.

The plot against $z_2$ (Fig. 5) shows that 13 of the 15 LEAs with the highest values of $z_2$ have positive residuals: four of these are in the top ten on the league table.

Examining plots of the residuals against the other background variables in the data set, we find a particularly striking non-linear trend in the plot against 16–18 population density (see Fig. 6). This suggests that the model parameters, the residuals and the rank order of the LEAs, would be substantially changed by inclusion of this variable. Such a finding is typical of analyses such as this one involving relatively few data points. It illustrates the difficulty in these analyses of choosing which variables to include and which to exclude.

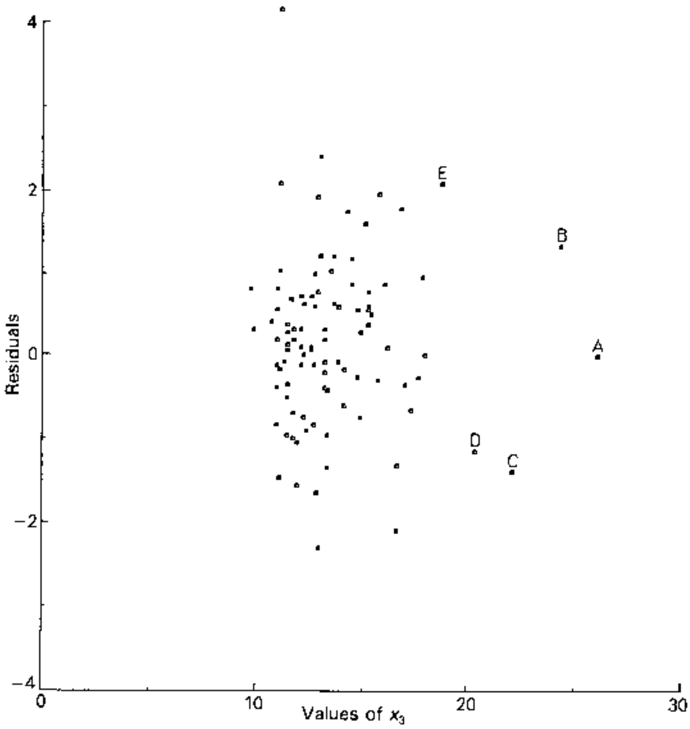

FIG. 3. Model 1 residuals plotted against $x_3$ (A=ILEA; B=Manchester; C=Haringey; D=Brent; E=Liverpool).

The results of this preliminary analysis of model 1 may be summarised as follows.

(1) The parameters of the model are poorly estimated.
(2) There is evidence of model mis-specification for which the data are not extensive enough for a proper investigation.

We next attempt to improve the fit and then go on to further examination of the stability of the residuals.

## IMPROVING THE FIT OF THE MODEL TO THE DATA

We do not propose to make major changes to the model. It is common practice to retain certain variables on substantive grounds even though their coefficients may be poorly estimated. Thus Gray & Jesson made a substantive case for including both $x_1$ and $x_2$ in their model and for excluding other variables 'irrelevant for policy purposes' which the DES had included. We therefore retain the four predictors $x_1$ to $x_4$ and for similar reasons retain the response variable $y$ as defined above and exclude 16–18 population density.

By limiting our changes to the model in this way we explore whether it is possible with this choice of variables:

(a) to develop a well-specified model which fits the data better than any other; or,

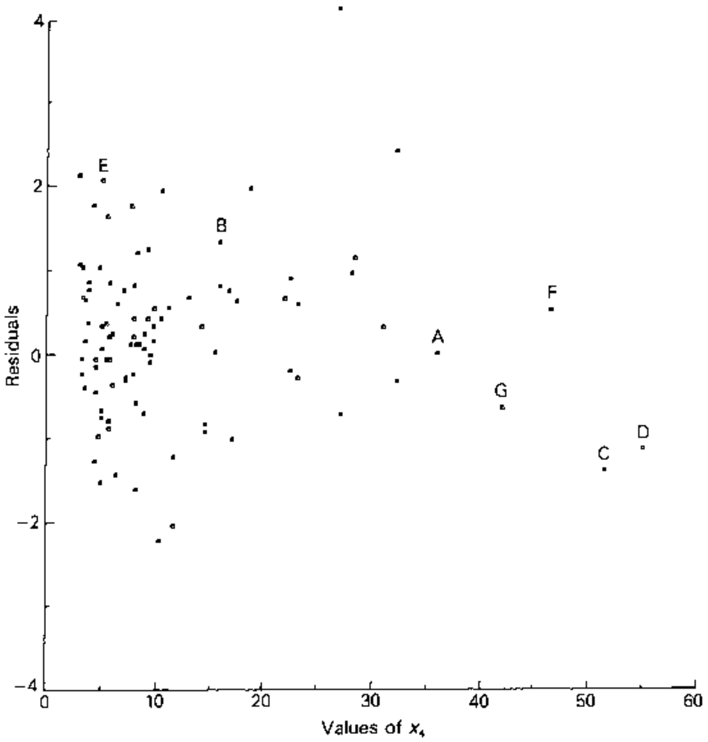

FIG. 4. Model 1 residuals plotted against $x_4$ (A=ILEA; B=Manchester; C=Haringey; D=Brent; E= Liverpool; F=Ealing; G=Newham).

(b) failing this, to place the LEAs in a stable rank order based on the residuals from any one of a number of candidate models.

Our conclusions show that:

(a) with only 96 data points it is possible to find different models using the same variables which fit the data equally well, with no objective way of choosing between them; and

(b) as models improve the rank order of their residuals becomes progressively more unstable. Thus two competing models fitting the data equally well may produce markedly different residual rankings. Indeed, this is to be expected, since as the model 'improves' so the residual variation approximates more closely to pure noise.

The dilemma this poses results directly from doing such an aggregate level analysis. Thus these limitations are inherent in the analysis and what follows is an illustration of their practical effects in a particular case. Similar conclusions would be expected to proceed from any other initial choice of variables.

In the model descriptions of this section and the next we use:

basic model

as an abbreviation for the right-hand side of model 1, i.e.

$$\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\beta_3 x_{3i}+\beta_4 x_{4i}+e_i$$

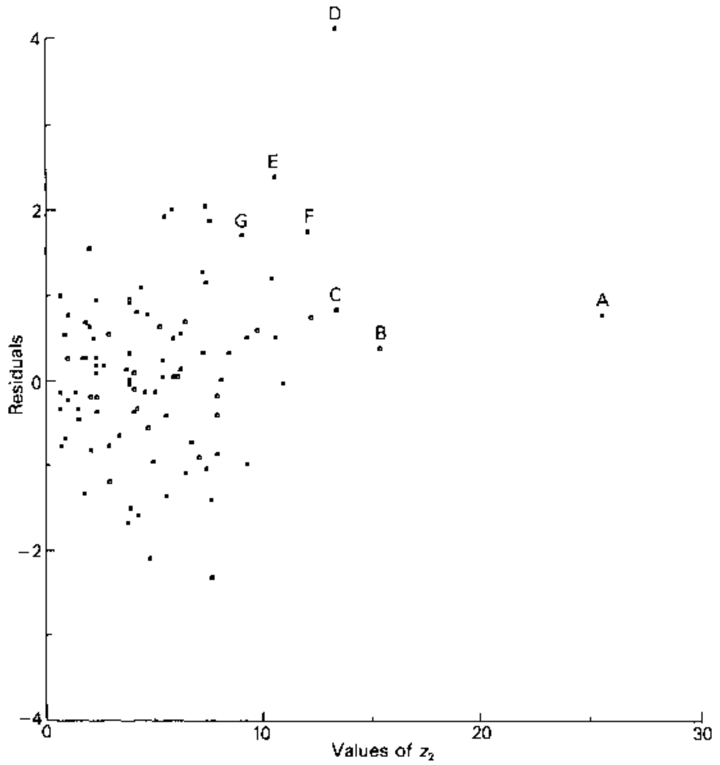

FIG. 5. Model 1 residuals plotted against $z_2$ (A=Richmond; B=Surrey; C=Croydon; D=Harrow; E=Barnet; F=Newcastle; G=Wirral).

We consider first the following alternative models:

$$y_i = \text{basic model} + \beta_{34} x_{3i} x_{4i} \qquad (2)$$

$$y_i = \text{basic model} + \beta_{11} x_{1i}^2 + \beta_{22} x_{2i}^2 \qquad (3)$$

$$y_i = \text{basic model} + \beta_{11} x_{1i}^2 + \beta_{22} x_{2i}^2 + \beta_{34} x_{3i} x_{4i} \qquad (4)$$

and then go on to explore routine transformations of the response variable.

Model 2 is suggested by the plot in Fig. 3 and model 3 by that in Fig. 2. Model 4 combines models 2 and 3.

The parameter estimates, etc., for each of these models are given in Table II.

For model 2, $R^2=0.84$. $\beta_2$ is again poorly estimated and the estimates of $\beta_2$, $\beta_3$ and $\beta_4$ have 'incorrect' signs. The plots of residuals against predicted scores, against $x_1$ and against $x_2$ continued to show a non-linear trend, as expected. The plot against $x_3$ (Fig. 7), however, shows that the addition of the product term has improved the fit of the model (compare Fig. 3).

The changes to the five authorities named in Fig. 7 are obvious and we should expect corresponding changes in their positions in the league table. The rest of the plot does not appear markedly different from Fig. 3. But concealed by this overall similarity are several other large changes in the positions of individual LEAs, some of which are shown in Table III.

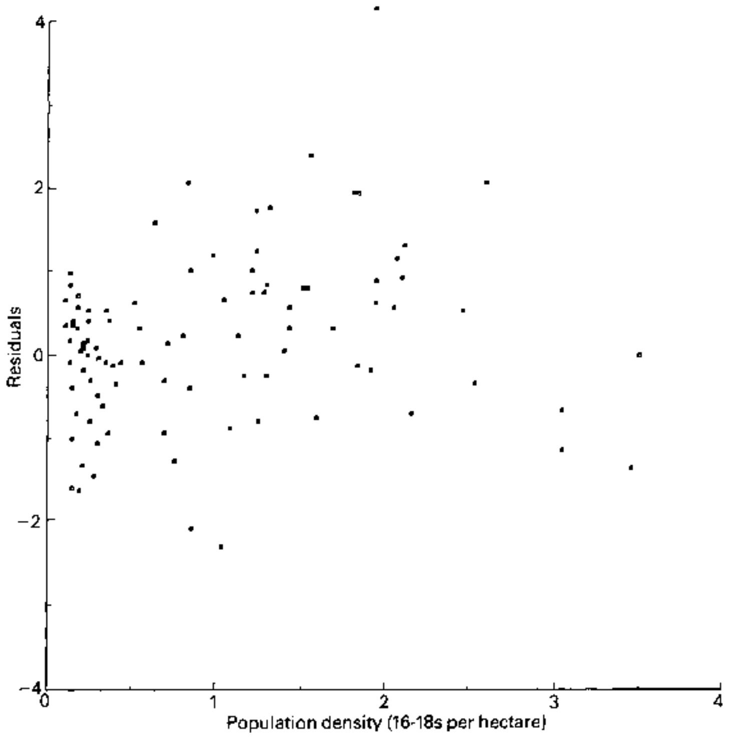

FIG. 6. Model 1 residuals plotted against 16–18 population density.

TABLE III. *Changes in the positions of selected LEAs resulting from adding quadratic and product terms in the predictors*

| LEA | Position under model 1 | Change under model 2 | Change under model 3 | Change under model 4 |
|---|---|---|---|---|
| ILEA[1,2] | 56 | up 45 | up 15 | up 44 |
| Haringey[1,2] | 91 | up 30 | up 5 | up 27 |
| Brent[1,2] | 88 | up 19 | up 7 | up 21 |
| Cheshire | 75 | up 11 | up 8 | up 13 |
| Northumberland | 40 | up 17 | up 2 | up 15 |
| Kingston | 22 | down 7 | down 23 | down 17 |
| Knowsley | 67 | down 12 | down 27 | down 26 |
| Kirklees | 29 | down 3 | down 11 | down 9 |
| Salford | 55 | down 16 | down 11 | down 17 |
| Ealing[2] | 34 | down 19 | up 6 | down 14 |
| Hounslow | 43 | down 16 | up 8 | down 13 |
| Havering | 46 | up 9 | down 16 | down 3 |
| Newham[2] | 77 | down 3 | down 12 | down 8 |
| Manchester[1] | 10 | down 10 | down 1 | down 9 |
| Liverpool[1] | 4 | down 3 | down 2 | down 4 |
| Harrow | 1 | no change | no change | no change |
| Bromley | 96 | no change | no change | no change |

[1] Indicates LEA with high value of $x_3$.

[2] Indicates LEA with high value of $x_4$.

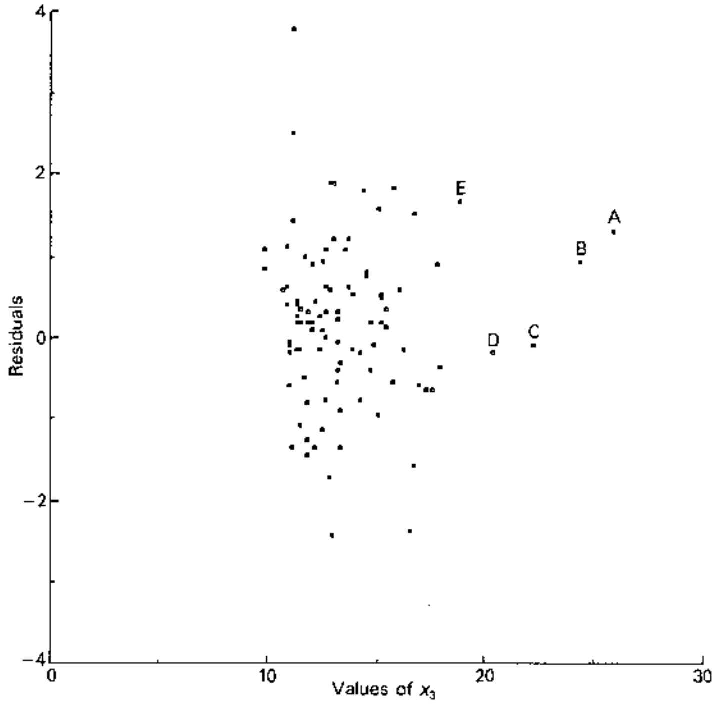

FIG. 7. Model 2 residuals plotted against $x_3$ (A=ILEA; B=Manchester; C=Haringey; D=Brent; E=Liverpool).

ILEA moves up 45 places from 56th to 11th, Haringey 30 places from 91st to 61st and Brent 19 places from 88th to 69th. Three other LEAs (Northumberland, Cheshire and Lincolnshire) move up 10 places or more. By contrast, Ealing drops 19 places from 34th to 53rd, Hounslow, Redbridge and Salford each drop 16 places and a further nine LEAs drop 10 or more places, including Manchester as expected. Thus our first heuristic attempt to improve the fit of the model has already had a considerable and widespread effect on the league table.

For model 3, $R^2=0.83$. The parameter estimates (see Table II) are difficult to interpret since the pairs $(b_1, b_{11})$ and $(b_2, b_{22})$ are highly negatively correlated ($r=-0.96$ and $-0.98$ respectively). The non-linear trend disappears, however, from the residual plot against predicted scores (see Fig. 8) and also from the plots against $x_1$ and $x_2$. The plot against $x_3$ was similar to that for model 1.

The inclusion of the extra terms again has a marked effect on the positions of some LEAs (see Table III) but different from model 2. The effects on ILEA, Brent and Haringey are less marked, but Knowsley this time drops 27 places to 94th and Kingston-upon-Thames 23 places to 45th. Seven other LEAs drop more than 10 places. Ealing and Hounslow, losers under model 2, move up instead 6 and 8 places, respectively. Havering, among the top 10 gainers under model 2, drops 16 places under model 3.

Adding the product term $\beta_{34}x_{3i}x_{4i}$ to model 3 to produce model 4 increases $R^2$ to 0.84. The values of $b_1$, $b_{11}$ and $b_2$ (see Table II) are little changed from model 3. $b_3$ and $b_4$ are, however, positive as for model 2. $b_4$ is estimated to be highly negatively correlated with $b_{34}$ ($r=-0.97$), while $r(b_3,b_{34})$ is estimated to be $-0.74$.

As expected, the plot of residuals against $x_3$ was similar to model 2. The plot against

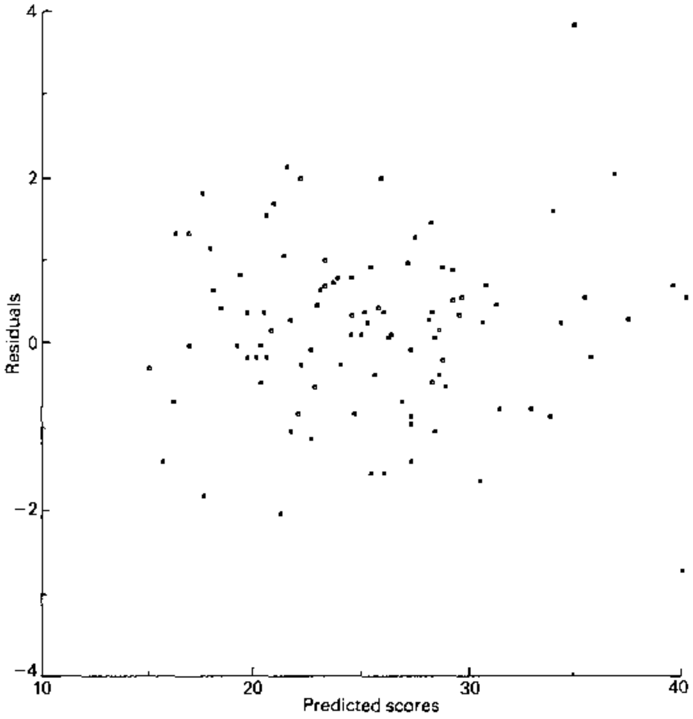

FIG. 8. Model 3 residuals plotted against predicted scores.

predicted scores was similar to model 3. As Table III shows, the changes in position between model 1 and model 4 were sometimes like model 2 and sometimes like model 3.

Models 2, 3 and 4 were all suggested by our preliminary analysis of model 1. To analyse such alternative models is part of the standard procedure for improving model fit. Allowing for these quadratic and cross-product effects changes the residuals and rankings, demonstrating instability while hardly changing the overall goodness of fit.

Another standard procedure is to transform the response variable. We first try the following transformations on their own:

square root: $\sqrt{y_i} = \text{basic model}$ (5)

logarithmic: $\log(y_i/100) = \text{basic model}$ (6)

logit: $\log(y_i/(100-y_i)) = \text{basic model}$ (7)

and

complementary log-log: $\log(-\log(1-y_i/100)) = \text{basic model}$ (8)

These transformations are commonly used with percentage or frequency data. In this case the response scale is arbitrary and there is no substantive theory to guide us, and so the choice of transformation is in fact much wider than this. The four above were similar to each other in their effects.

In all four cases the parameters $\beta_2$ and $\beta_3$ remain poorly estimated (see Table IV).

The three residual plots, against predicted scores, $x_1$ and $x_2$, showed no trends. Those against $x_3$ and $x_4$ remained similar to those for model 1. Thus models 5 to 8 were similar to model 3 in their effects on the overall residual distribution.

TABLE IV. *Parameter estimates and $R^2$ for square root, logarithmic, logit and log-log transformations of the response variable*

| | Model 5<br>(square root) | Model 6<br>(logarithmic) | Model 7<br>(logit) | Model 8<br>(log-log) |
|---|---|---|---|---|
| $b_0$ | 2.85 (0.45) | −2.25 (0.9) | −2.26 (0.25) | −2.25 (0.22) |
| $b_1$ | 0.07 (0.008) | 0.03 (0.003) | 0.04 (0.004) | 0.03 (0.004) |
| $b_2$ | 0.006 (0.014) | 0.002 (0.006) | 0.003 (0.008) | 0.003 (0.007) |
| $b_3$ | 0.004 (0.014) | 0.001 (0.006) | 0.002 (0.008) | 0.002 (0.007) |
| $b_4$ | −0.004 (0.003) | −0.002 (0.001) | −0.002 (0.002) | −0.002 (0.001) |
| $s^2$ | 0.07 | 0.01 | 0.02 | 0.02 |
| $R^2$ | 0.83 | 0.81 | 0.82 | 0.82 |

As regards changes in the rankings of the residuals (see Table V for the most extreme) the top gainers are the same set of LEAs, with minor variations in order, for all four transformations as are the top losers. All four transformations produce some individual positional losses which are more pronounced than any individual gains.

TABLE V. *Most extreme changes in position for square root, logarithmic, logit and log-log transformations*

| LEA | Position under model 1 | Change under model 5 | Change under model 6 | Change under model 7 | Change under model 8 |
|---|---|---|---|---|---|
| Ealing | 34 | up 13 | up 21 | up 17 | up 18 |
| Hounslow | 43 | up 11 | up 13 | up 12 | up 12 |
| ILEA | 21 | up 8 | up 15 | up 11 | up 11 |
| Leicestershire | 57 | up 5 | up 12 | up 10 | up 11 |
| Brent | 88 | up 4 | up 13 | up 9 | up 10 |
| Newham | 77 | down 15 | down 18 | down 17 | down 17 |
| Sandwell | 66 | down 12 | down 23 | down 19 | down 20 |
| Kingston | 22 | down 11 | down 27 | down 15 | down 22 |
| Knowsley | 67 | down 19 | down 25 | down 21 | down 24 |
| Richmond | 23 | down 21 | down 40 | down 29 | down 34 |
| Surrey | 36 | down 25 | down 40 | down 31 | down 36 |

We select model 7 (the logit transformation) for further consideration. As this model has apparently similar effects to model 3 on the overall distribution of residuals it is natural to ask whether the positional changes also are similar. Table VI shows that they are not.

Finally, we add the term $\beta_{34}x_{3i}x_{4i}$ to model 7:

$$\text{logit}\ (y_i) = \text{basic model} + \beta_{34}x_{3i}x_{4i} \tag{9}$$

When this model is fitted to the data, $R^2$ increases somewhat to 0.83. The parameter estimates, with their standard errors, are:

| | | |
|---|---|---|
| $b_0$ | −2.39 | (0.25) |
| $b_1$ | 0.04 | (0.004) |
| $b_2$ | 0.0008 | (0.008) |
| $b_3$ | 0.02 | (0.01) |
| $b_4$ | 0.01 | (0.006) |
| $b_{34}$ | −0.0008 | (0.0004) |
| $s^2$ | 0.02 | |

TABLE VI. *Models with similar overall fit compared for their effects on selected LEAs*

| LEA | Position under model 1 | Change under model 3 | Change under model 7 | Change under model 4 | Change under model 9 |
|---|---|---|---|---|---|
| Surrey | 36 | up 4 | down 31 | up 10 | down 29 |
| Richmond | 23 | down 4 | down 29 | down 7 | down 35 |
| Barking | 79 | up 2 | down 13 | down 3 | down 14 |
| Buckinghamshire | 33 | up 2 | down 13 | up 5 | down 11 |
| Sandwell | 66 | down 4 | down 19 | down 11 | down 22 |
| Barnsley | 61 | up 3 | down 12 | up 8 | down 13 |
| West Sussex | 47 | down 3 | down 11 | up 7 | down 8 |
| Rotherham | 41 | up 4 | down 9 | up 9 | down 4 |
| Kingston | 22 | down 23 | down 15 | down 17 | down 26 |
| Ealing | 34 | up 6 | up 17 | down 14 | up 6 |
| Merton | 32 | down 12 | up 3 | down 22 | down 5 |
| Bexley | 63 | down 11 | up 6 | down 6 | up 6 |
| Devon | 39 | down 14 | up 4 | down 13 | up 1 |
| Kirklees | 29 | down 11 | up 7 | down 9 | up 2 |
| Havering | 46 | down 16 | up 3 | down 3 | up 11 |
| Hounslow | 43 | up 8 | up 12 | down 13 | up 3 |

*Note:* Models 3 and 7 showed a similar overall fit to the data, as did models 4 and 9.

The values of $b_0$ and $b_1$ are little changed from model 7 (see Table IV). $\beta_2$ is again poorly estimated and close to zero. $b_4$ is estimated to be highly negatively correlated with $b_{34}$ ($r=-0.96$), whereas $r(b_3, b_{34})$ is estimated to be $-0.68$.

As with model 4, the addition of the product term improved the residual plot against $x_3$: all the diagnostic plots for model 9 in fact were similar to the corresponding plots for model 4. Yet despite this similarity Table VI shows that the league tables produced by models 4 and 9 are dissimilar.

We may summarise the results of trying to improve the fit of the basic model as follows.

(1) It is possible to improve the fit in predictable ways by following standard procedures suggested by the residual plots for the basic model.

(2) In every model several of the parameters are found to be poorly estimated.

(3) No model can claim to be 'the best' for the whole data space.

(4) Models 2 to 9 all produce marked changes in the league positions of some individual LEAs when compared with the basic model.

(5) General plots of residuals for diagnosis of the models often conceal large changes in the positions of individual LEAs. Thus different models which apparently fit the data similarly are capable of producing quite different league tables.

(6) The residual rankings become more, not less, unstable as the fit to the data improves.

## STABILITY OF THE RESIDUALS: FURTHER EXAMINATION

Throughout the previous section we followed standard procedures for improving the fit of the model to the data. It may be thought that the instability of the residuals that we found has already amply demonstrated their unsuitability as measures of efficiency. It did appear, however, that certain LEAs were relatively stable in position. In particular, the top ten, apart from Manchester, stayed unchanged as a set, the LEAs merely swapping places with each other, if indeed moving at all. Perhaps these LEAs can fairly be described as more efficient than the rest at getting their students five or more O-level equivalents, after adjusting for the five variables in the original model. To claim this is certainly less than to claim to rank the LEAs from first to ninety-sixth.

We now show how this more modest claim may be tested by adding further quadratic and product terms to model 7 and examining the effects on the model parameters and the residual rankings. Table VII gives the parameter estimates, and $R^2$, for the following models, with model 7 for comparison:

$$\begin{aligned}\text{logit}\,(y_i) = {} & \text{basic model}\\ & +\beta_{23}x_{2i}x_{3i}+\beta_{24}x_{2i}x_{4i}+\beta_{34}x_{3i}x_{4i}\end{aligned} \tag{10}$$

$$\begin{aligned}\text{logit}\,(y_i) = {} & \text{basic model}\\ & +\beta_{11}x_{1i}{}^2+\beta_{22}x_{2i}{}^2+\beta_{33}x_{3i}{}^2+\beta_{44}x_{4i}{}^2\end{aligned} \tag{11}$$

and

$$\begin{aligned}\text{logit}\,(y_i) = {} & \text{basic model}\\ & +\beta_{11}x_{1i}{}^2+\beta_{22}x_{2i}{}^2+\beta_{33}x_{3i}{}^2+\beta_{44}x_{4i}{}^2\\ & +\beta_{23}x_{2i}x_{3i}+\beta_{24}x_{2i}x_{4i}+\beta_{34}x_{3i}x_{4i}\end{aligned} \tag{12}$$

Table VIII shows the positional changes for the top and bottom ten LEAs under model 1 when models 10, 11 and 12 are fitted. The positions of Barnet and Manchester are now shown to be highly unstable in addition to those of Brent and Haringey. Harrow moves off the top for the first time.

TABLE VII. *Parameter estimates and $R^2$ for models 7, 10, 11 and 12*

| | Model 7 | Model 10 | Model 11 | Model 12 |
|---|---|---|---|---|
| $b_0$ | −2.3 (0.25) | −2.3 (0.48) | −2.12 (0.48) | −2.4 (0.49) |
| $b_1$ | 0.04 (0.004) | 0.04 (0.004) | 0.06 (0.02) | 0.06 (0.01) |
| $b_2$ | 0.003 (0.008) | 0.003 (0.02) | −0.04 (0.04) | −0.07 (0.04) |
| $b_3$ | 0.002 (0.008) | −0.005 (0.03) | −0.02 (0.04) | −0.07 (0.04) |
| $b_4$ | −0.002 (0.002) | 0.02 (0.007) | 0.004 (0.004) | 0.03 (0.009) |
| $b_{11}$ | | | −0.0005 (0.0003) | −0.0005 (0.0003) |
| $b_{22}$ | | | 0.001 (0.0009) | 0.001 (0.001) |
| $b_{33}$ | | | 0.006 (0.0001) | 0.005 (0.003) |
| $b_{44}$ | | | −0.0002 (0.0001) | 0.0002 (0.0002) |
| $b_{23}$ | | 0.0006 (0.001) | | −0.003 (0.003) |
| $b_{24}$ | | −0.0005 (0.0004) | | −0.002 (0.001) |
| $b_{34}$ | | −0.0007 (0.0003) | | −0.0004 (0.0004) |
| $s^2$ | 0.02 | 0.02 | 0.02 | 0.02 |
| $R^2$ | 0.82 | 0.84 | 0.83 | 0.85 |

We could continue, for example, by exploring further transformations of the response variable, but it is already clear that positions apparently near the top or the bottom of the league according to some models are in fact unstable. Certain LEAs such as Haringey and Brent appear sensitive to all model variations of a particular type, suggesting that for these LEAs a particular factor might be significant (but which the data are in any case inadequate to explore). Others such as Barnet and Manchester show sudden unpredictable movement. No LEA (not even Harrow) is completely unaffected.

TABLE VIII. *Positional changes to the top and bottom ten under models 10, 11 and 12*

| LEA | Position under model 1 | Change under model 10 | Change under model 11 | Change under model 12 |
|---|---|---|---|---|
| Harrow | 1 | no change | no change | down 3 |
| Barnet | 2 | down 33 | down 4 | down 25 |
| St Helens | 3 | up 1 | up 1 | up 1 |
| Liverpool | 4 | down 3 | up 1 | down 1 |
| Coventry | 5 | up 2 | down 2 | up 4 |
| Sutton | 6 | no change | down 6 | down 1 |
| Newcastle | 7 | up 3 | up 3 | up 1 |
| Wirral | 8 | up 3 | up 3 | no change |
| Cleveland | 9 | up 1 | up 1 | no change |
| Manchester | 10 | down 4 | down 6 | down 47 |
| Bedfordshire | 87 | up 1 | up 1 | down 2 |
| Brent | 88 | up 16 | up 33 | up 19 |
| Rochdale | 89 | down 2 | down 1 | down 1 |
| Isle of Wight | 90 | up 6 | up 2 | up 3 |
| Haringey | 91 | up 12 | up 12 | up 25 |
| Essex | 92 | up 3 | up 3 | up 6 |
| Norfolk | 93 | up 3 | up 2 | up 2 |
| Northamptonshire | 94 | up 2 | up 2 | up 2 |
| Oldham | 95 | down 1 | down 1 | down 1 |
| Bromley | 96 | up 2 | up 1 | up 2 |

## CONCLUSIONS

We have demonstrated that the aggregate-level models used for LEA comparisons suffer severe problems of interpretation. Small changes in the input variables, in particular the inclusion of non-linear terms, change the rank ordering of the regression residuals. Non-linear transformation of the response variable likewise changes rankings. The usual overall measures of model 'fit', such as multiple correlation and residual variance, are not affected in the same way and reliance solely on these is therefore misleading.

In typical aggregate-level analyses the total population studied is finite and often rather small. In the present case there are only 96 LEAs. If we wished to generalise from such an analysis we would have to invoke a 'superpopulation' model where the obtained data were considered to be a random sample from a conceptually infinite possible population. We would need to take into account the possibility of change over time and to ask whether any relationship observed could be assumed to hold for the future. The difficulty here is that variables such as LEA spending have a complex relationship with other explanatory variables such as socio-economic composition, and changes in the values of any of these variables could be expected to change the relationships. Thus, at the very least, a replication of such analyses is needed to study time trends.

The relatively small sample size presents technical problems also, in that important predictors may not be detected because of high 'sampling variability'. We would expect to find different combinations of explanatory variables which provide equally good 'fits' with no objective way of choosing between them. These kinds of analyses thus contain inherent indeterminacies which do not allow us objectively to choose any one version with its associated rank ordering.

We have also alluded to the residual scaling issue. It is common to make comparisons based upon the actual residuals from the analysis. Adjusting for differences in their distribution by standardising the residuals arguably produces a fairer rank order for a given model: we have shown, however, that it does not produce stability against small changes to the model.

It is clear, therefore, that the usual procedure for ranking LEAs using aggregate data has little justification. As has been pointed out elsewhere (Aitkin & Longford, 1986; Goldstein, 1987) the only secure basis for attempting to compare schools or LEAs is to use a proper multilevel modelling procedure which simultaneously measures student characteristics as well as those of schools or LEAs. By incorporating variables which measure such characteristics we can hope to gain a deeper understanding of the factors which influence school performance.

The analyses in this paper have important implications for attempts to compare schools and LEAs using either exam results or the results of large-scale national assessment and testing programmes.

## REFERENCES

AITKIN, M. & LONGFORD, N. (1986) Statistical modelling issues in school effectiveness studies, *Journal of the Royal Statistical Society A*, 149(1), pp. 1–43.

CHARNES, A., COOPER, W.W. & RHODES, E. (1978) Measuring the efficiency of decision making units, *European Journal of Operational Research*, pp. 429–444.

DES (1983) School standards and spending: statistical analysis, *Statistical Bulletin 16/83* (London, DES).

DES (1984) School standards and spending: statistical analysis: a further appreciation, *Statistical Bulletin 13/84* (London, DES).

GOLDSTEIN, H. (1987) *Multilevel Models in Educational and Social Research* (London, Griffin).

GRAY, J. & JESSON, D. (1987) Exam results and local authority league tables, *Education and Training U.K. 1987*, pp. 33–41.

ILEA (1980) *School Examination Results in the ILEA 1978* (London, ILEA).

JESSON, D., MAYSTON, D. & SMITH, P. (1987) Performance assessment in the education sector: educational and economic perspectives, *Oxford Review of Education*, 13, pp. 249–266.

LEVITT, M. & JOYCE, L. (1987) *The Growth and Efficiency of Public Spending*, pp. 107–121 (Cambridge, Cambridge University Press).

*Correspondence:* Geoffrey Woodhouse & Professor Harvey Goldstein, Institute of Education, University of London, 20 Bedford Way, London WC1H 0AL, United Kingdom.

## APPENDIX

### *Introduction*

Data Envelopment Analysis (DEA) arose from attempts to define measures of 'efficiency' which could be applied to non-commercial enterprises such as schools and hospitals. In essence, efficiency is defined as a weighted sum of 'outputs' divided by a (differently) weighted sum of inputs, with the weights estimated from the data. Each unit, such as a school or a hospital, is assumed to have a measurement on each output and each input. In the case of schools the outputs might consist, for example, of average examination result or attendance rate over a period of time, with inputs such as expenditure per pupil or average intake test score.

A distinction is immediately apparent from the class of models, based on linear regression, which seek to relate an output measure to a set of inputs and then interpret residuals as measures of efficiency. In DEA there is no attempt to take account of the nature of any relationships between output and input, rather it proceeds to *define* efficiency as a ratio. It does, however, use data aggregated to the level of the unit as do the analyses described in the present paper.

In the following sections we outline the DEA procedure and discuss its interpretation using simple examples.

### *The DEA Model*

A formal statement of this model is as follows.

We denote the rth output of unit j by $y_{rj}$ and the ith input of unit j by $x_{ij}$. For the jth unit we shall require weights $u_{rj}$, $v_{ij}$ as follows.

There are p units. For the jth unit, define the kth ratio:

$$h_{jk}=\frac{\Sigma_r u_{rj} y_{rk}}{\Sigma_i v_{ij} x_{ik}} \qquad (k=1,\ldots,p) \tag{1}$$

The required weights are those which make this ratio, for k=j, as large as possible relative to all other units. Before we can do this, however, we need to 'constrain' the solution. As it stands we can multiply the weights in the numerator all by any number, say *a*, and the weights in the denominator by any number, say *b*, without altering the nature of the solution. There are many ways to 'fix' the weights, for example by requiring each set to add to a constant, say 1.0. The approach of DEA is to require all the ratios to be no greater than 1.0, with a further constraint on the scale of either the numerator or denominator weights; one choice again being that either should sum to 1.0. We note that the solution will generally be dependent on the constraints chosen, in a non-trivial way, and this appears to be a problem which has not been studied in this context. We shall not pursue this issue, however, since there are more serious concerns.

The above procedure is carried out for each unit in turn, thus giving a set of coefficient values and a value of the ratio for each unit. These ratios are then interpreted as the relative 'efficiencies' of each unit.

To simplify matters consider the common case where there is only one output, say a school's average examination result, with several inputs. The procedure now seeks to maximise:

$$h_{jk}=\frac{y_k}{\Sigma_i v_{ij} x_{jk}} \tag{2}$$

The weights and the input and output variables are all assumed to be positive so that consistent interpretations are possible.

We can further simplify (2) by considering the special case of just one input variable. We no longer now have a complex maximising problem, since we simply compare the ratios:

$$h_k = \frac{y_k}{x_k} \tag{3}$$

for each unit. We can gain some insight into the procedure by studying this case in more detail.

### *Comparing Simple Ratios*

In model (3) let us suppose that our output is an average examination result and our input is an average intake score for a school. Then the ratio of these two variables is defined as efficiency and a simple ranking of schools can be made. The interpretation, however, is not straightforward.

First, while it could be argued that the ratio of test or exam scores is a reasonable measure of efficiency, it would be more difficult to argue that a simple ratio of an exam score to the proportion of children from middle-class homes is a good measure of efficiency. In DEA analyses with multiple input variables, these are typically a mixture of test scores and other characteristics of the school and the students attending it, so that some care needs to be paid to any interpretations.

Secondly, unlike statistical linear models, DEA pays no attention to a description of the actual relationship between output and input variables, simply concerning itself with the properties of the ratio. Nevertheless, a particular relationship still exists, and the nature of this relationship will determine the properties of the ratio. In our simple case let us suppose that the true relationship has the following form:

$$y_k = a + bx_k \tag{4}$$

Then we obtain immediately:

$$h_k = \frac{a}{x_k} + b \tag{5}$$

and we see that the ratios are inversely proportional to the input scores, or intake test scores in our example. In this case it would be quite misleading to interpret the $h_k$ as 'efficiencies'. A similar argument applies to the case of multiple inputs where DEA effectively maximises a ratio of two linear functions of the input variables.

Thus, even in the simplest cases we see that the DEA model used on its own will not have an unambiguous interpretation. Great care is needed in defining input (and output) variables so that proper interpretations of the term 'efficiency' can be made, and more importantly, it is clear that some information about the *relationship* between output and input variables is required. This latter requirement leads to the use of standard statistical models.

Needless to say, in the more general case with multiple inputs and outputs the same objections apply, although the complexity of the estimation procedures necessary may sometimes obscure this point. It is also worth pointing out that while the use of DEA to judge such issues as school effectiveness is highly problematical, in more limited contexts it may be useful. Thus, if we are concerned to measure the efficiency of several industrial processes with monetary inputs and outputs, a definition of efficiency as a ratio based on costs might well be a useful basis for comparison.

## *Conclusion*

While there is a certain attraction in using a simple ratio of multiple inputs and outputs, it can lead to serious oversimplifications. At the very least, as shown in the present paper, extensive sensitivity analyses are required, and experience with these suggests that stable interpretations in terms of efficiency or effectiveness are very difficult, if not impossible, to make.

## References

**Performance Assessment in the Education Sector: Educational and Economic Perspectives**
David Jesson; David Mayston; Peter Smith
*Oxford Review of Education*, Vol. 13, No. 3. (1987), pp. 249-266.
Stable URL:
http://links.jstor.org/sici?sici=0305-4985%281987%2913%3A3%3C249%3APAITES%3E2.0.CO%3B2-F