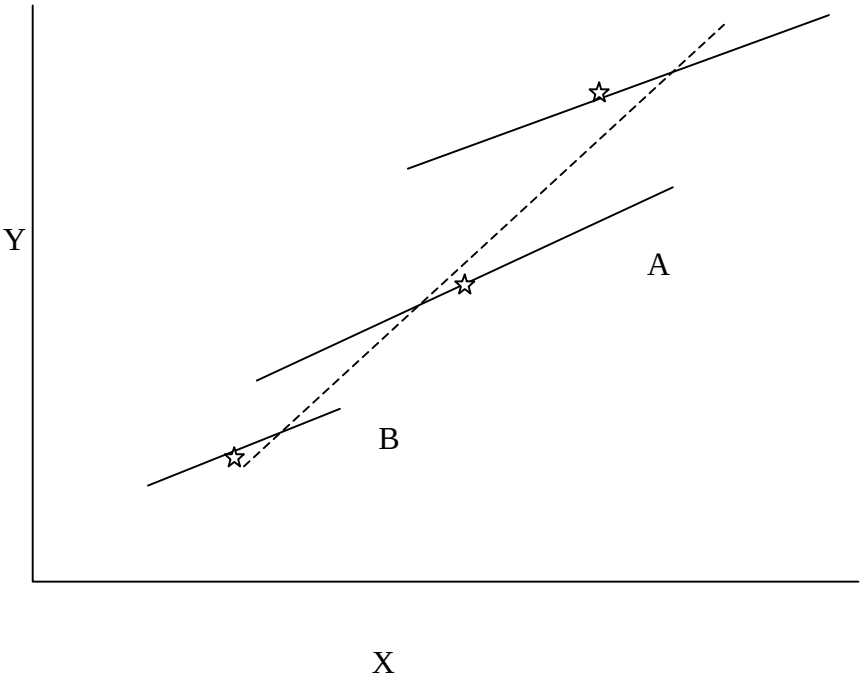
Interpreting aggregate level models

A simple 2-level example

Consider the following illustration of a series of linear relationships for a set of higher level units. For convenience assume that students are level 1 units and schools level 2 units and that the response and predictor variables are test scores. Assume for simplicity also that all random variables are Normally distributed.

Figure 1



Consider first a model which relates the mean school Y value to the mean school X value as indicated by the symbols, and suppose this is linear as shown by the dotted line in figure 1.

We write

$$y_{.j} = \alpha_0 + \beta_0 x_{.j} + v_{0j}, \quad v_{oj} \sim N(0, \sigma^2_{v0}) \tag{1}$$

where $y_{.j}$, $x_{.j}$ are population means for the *j*-th school.

We can now write a model for the student responses as

$$y_{ij} - y_{.j} = \beta(x_{ij} - x_{.j}) + e_{ij}, \quad \mathrm{e_{ij}} \sim N(0, \sigma^2_{e0}) \tag{2}$$

and on substitution

$$y_{ij} = \alpha_0 + \beta^* x_{.j} + \beta x_{ij} + v_{0j} + e_{ij}$$
$$\beta^* = \beta_0 - \beta \tag{3}$$

To complete the multilevel formulation we assume that the between-school differences are randomly distributed and write

$$\beta * x_{.j} = v_{1j} \sim N(0, \sigma^2_{v1}) \qquad (4)$$

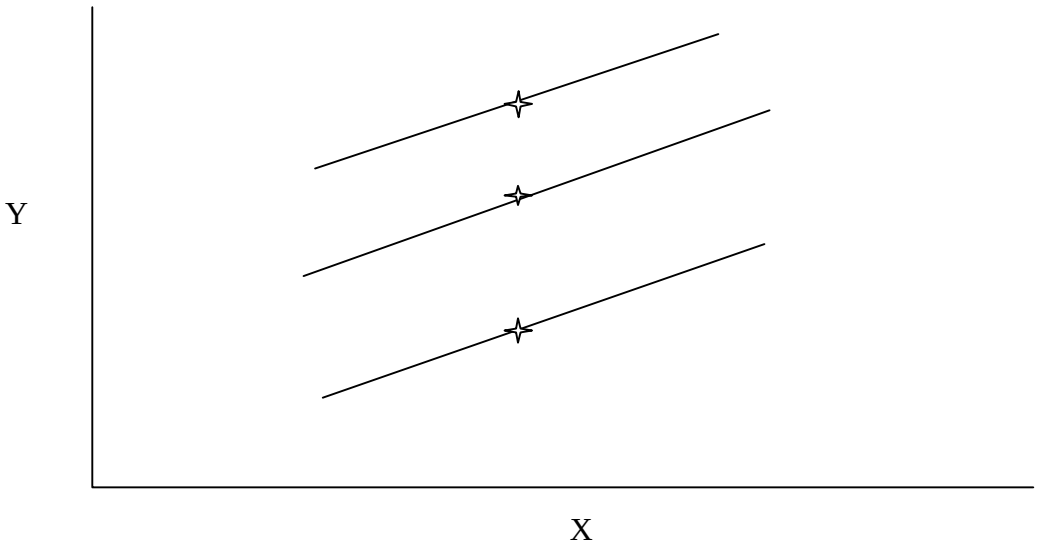
where, without loss of generality we may assume $x_{..} = 0$. This gives

$$\begin{aligned} y_{ij} &= \alpha + \beta x_{ij} + u_{0j} + e_{0ij} \\ u_{0j} &= v_{0j} + v_{1j} \qquad (5) \\ \alpha &= \alpha_0 + \mu_{\beta *} \end{aligned}$$

which is a standard 2-level model. In this model the school effects are the estimates $\hat{u}_{0j}$. From the aggregate level analysis, however, we are estimating v_{0j} which is not what is required, and simply represents the extent to which the school means do not lie on a straight line, or any other smooth curve fitted to the data.

If the $x_{.j}$ are small however ($x_{..} = 0$), so that the v_{1j} are small compared to the v_{0j}, then the aggregate and individual level analyses will give approximately the same results. In the extreme case where the $x_{.j}$ are zero we have the following data structure:

Figure 2



The individual level model is

$$y_{ij} = \alpha + \beta x_{ij} + u_{0j} + e_{0ij}$$

averaging over the population of (X,Y) values we obtain

$$y_{.j} = \alpha + \beta x_{.j} + u_{0j}$$
$$y_{.j} = \alpha + u_{0j}, \quad x_{.j} = x_{..} = 0$$

and the $y_{.j} - y_{..}$ will provide the required estimates. Thus, inferences from an aggregate level analysis will be correct if schools with the same mean X value are compared. In some circumstances we may be able to approximate this by making comparisons within narrow ranges of X values, but only if the distributions within such ranges are the same so that the X means are equal. This will generally not be the case if, overall, schools have very different means.

Finally, to illustrate the issues, consider Figure 1 again. Here we see that, using student level information those in school B have lower mean Y value than those in school A at least for any given X value within the range of overlap. Using the aggregate level analysis this inference is actually reversed.

More complex models

So far we have dealt with simple variance component models only. In figure 1, if the lines are no longer parallel, so that we have a random coefficient model, then we obtain similar results, using an extension of the previous notation, where (3) and (4) become

$$\begin{aligned} y_{ij} &= \alpha_0 + \beta_j * x_{.j} + \beta_j x_{ij} + v_{0j} + e_{0ij} \\ \beta_j * &= \beta_0 - \beta_j \\ \beta_j * \; x_{.j} &= v_{1j} \sim N(\mu_{\beta*}, \sigma^2_{v_1}) \end{aligned} \qquad (6)$$

etc.

If our model becomes more complex, however, say involving an interaction between two level 1 explanatory variables as follows

$$\begin{aligned} y_{ij} &= \alpha + \beta x_{ij} + \gamma z_{ij} + \delta x_{ij} z_{ij} + u_{0j} + e_{0ij} \\ y_{.j} &= \alpha + \beta x_i + \gamma z_i + \delta \left(\overline{x_{ij} z_{ij}} \right) + u_{0j} + e_{0.j} \end{aligned} \qquad (7)$$

then clearly, since $x_{.j} z_{.j} \neq \overline{x_{ij} z_{ij}}$, we cannot estimate the parameters of the aggregate model in the second line of 5 without access to the level 1 data. The same problem arises if the individual level model includes polynomial terms. Note that (7) implies that the relationship between Y and X depends on Z, where Z may be, for example, gender, social class, etc.

An example

To illustrate the issue we use a data set consisting of GCSE scores in 65 schools for 4059 pupils where a reading test score is available for all students at the age of 11 at entry to secondary school. The data are described in Goldstein et al (1998).

The following model is fitted to the student level data

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + u_j + e_{ij} \qquad (8)$$

where the variables x_{hij}, $h = 1,...4$ are respectively a reading test score at 11 years, gender, belonging to the top 25% on a verbal reasoning test at 11 years and belonging to the middle 50% on the verbal reasoning test at 11 years. All variables are standardised. The data and a detailed analysis are described in Goldstein et al. (1998). The following is the aggregated model fitted to the same data

$$y_{.j} = \beta_0^* + \beta_1^* x_{1.j} + \beta_2^* x_{2.j} + \beta_3^* x_{3.j} + \beta_4^* x_{4.j} + u_j^* \qquad (9)$$

Figure 3 shows the residuals u_j^* from (9) plotted against the u_j from (8).

Figure 3 Model comparison fitting all predictors. r=0.88

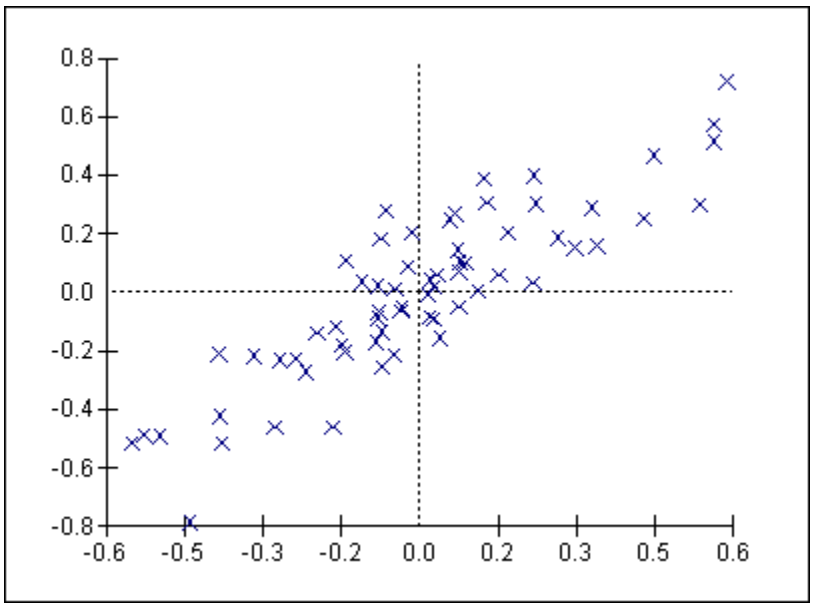
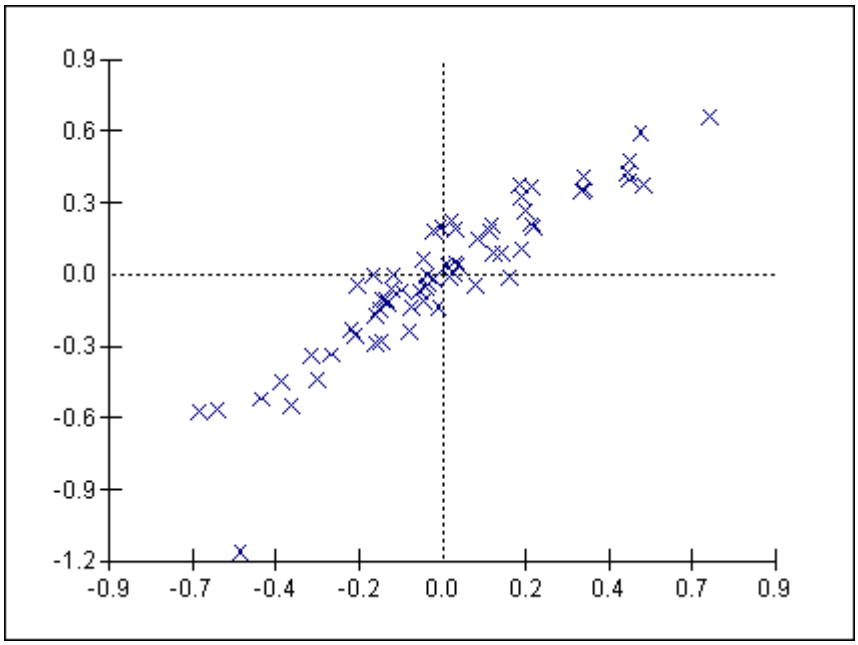


Figure 4 shows the residuals from a similar pair of analyses which fit just an intercept and the reading test score.

Figure 4 Model comparison fitting intercept and reading test score. r=0.92



It is clear from Figure 3 and Figure 4 that many schools alter their relative positions, by as much as 20% in the rankings. In effect, Figure 4, which represents the simpler analysis, is the case where, compared to Figure 3 which includes more explanatory variables, the X values have a smaller spread, so that we would expect the student level and aggregate level analyses to be more alike, as is the case.

Conclusions

It is clear from this exposition, that aggregate level analyses which fit models using average values of predictor and response variables do not in general provide appropriate inferences about school effects since they are estimating different components of any underlying student level model. Moreover, as they become more misleading as further relevant variables are included in the model.

Essentially, residuals from the aggregate analysis (1) are simply estimating the extent to which the school means depart from a smooth curve relating the Y and X values. These residuals have no necessary relationship with the residuals derived from a student level analysis which are estimating the differences between the means themselves, given X. If the degree or type of smoothing is changed the residuals and their rankings will also change, as demonstrated above and also by Woodhouse and Goldstein (1989).

Harvey Goldstein

9-Dec-98

References

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., et al. (1998). *A user's guide to MLwiN* . London, Institute of Education:

Woodhouse, G. and Goldstein, H. (1989). Educational Performance Indicators and LEA league tables. *Oxford Review of Education* **14**: 301-319.