Consistent Parameter Estimation for Lagged Multilevel Models

Neil H. Spencer
University of Hertfordshire

Abstract: The estimation of parameters of lagged multilevel models is considered. This type of model is used in many application areas, including psychology and education, where changes in test results over time can be modelled. Standard estimation techniques are shown to give inconsistent results for this formulation of the multilevel model. For two simple assumptions concerning the nature of the model covariate a first and second difference instrument methods for consistent estimation are developed. Simulations are used to demonstrate their success in obtaining consistent parameter estimates. Use of the instrument methods with more complex multilevel models is considered.

Key words: Consistent estimation, educational research, instrumental variables, longitudinal data, multilevel models.

1. Introduction

Repeated measures data are common in many areas of application including cognitive growth, educational testing, medical research and environmental pollution. The number of occasions for which data are available can range from just 2 (e.g. Goldstein and Thomas, 1996) to sometimes considerably larger numbers (e.g. 8 in Williamson et al, 1991; 12 in Rutter and Elashoff, 1994; 416 in Chock et al, 1975).

Multilevel modelling is often seen as an appropriate strategy to use when the data to be analysed has a hierarchical structure. This is the case when, for instance, students are grouped in classes which are grouped

within schools. A modelling strategy which does not allow for these groupings effectively assumes that all the students are independent of each other. This is evidently not a safe assumption to make as there may be factors that influence the response variable being measured that work at the class or school level. For example, a test mark obtained by a student in one school may be considered independent of that obtained by a student in a different school. However, test marks obtained by two students in the same class cannot be assumed to be independent of each other as both students have experienced the same standard of teaching, the same group dynamics in the class, etc. Indeed, test marks obtained by two students in different classes but in the same school cannot be assumed to be independent of each other because both students have experienced the same school resources, school atmosphere, etc. Proceeding with a standard regression analysis under the false assumption of independent observations leads to standard errors for estimates that are too small, giving false impressions of the importance of regressors. This also has a major potential impact on model building.

Random effects modelling, and more generally, multilevel modelling allows for the grouping in the data by means of partitioning the error according to the different levels that exist in the data. There are a number of books devoted to the subject of multilevel modelling (e.g. Bryk and Raudenbush (1992), Goldstein (1995), Hox (2002), Kreft and de Leeuw (1998), Longford (1993), Snijders and Bosker (1999). Good articles introducing multilevel modelling have been written by Paterson (1991), Paterson and Goldstein (1991), Gilthorpe et al (2000).

The commonplace use of multilevel modelling, however, did not start until the 1980s and the development of algorithms that could be applied to a wide range of formulations of the multilevel model, e.g. Mason et al (1984), Goldstein (1986), Raudenbush and Bryk (1986), Longford (1987). Computer packages that implement these algorithms (e.g. HLM, MLwiN, VARCL) are now widely used to model hierarchically structured data.

In the field of multilevel modelling, repeated measures data are usually modelled in one of two ways: (i) using growth curves and (ii) creating models where data collected at earlier measurement occasions are used as covariates (see Plewis, 1996, for a review). Examples of analyses employing growth curves include Williamson et al (1991), Rutter and Elashoff (1994).

The use of data from previous occasions as covariates in models is also widespread. Examples where the covariate is a pre-test variable or some previous measurement include papers by Aitkin et al (1981), Aitkin and Longford (1986), Goldstein and Thomas (1996). A further set of examples include papers by Milionis and Davies (1994) and Plewis (1991). These two sets of examples differ in that the dependent variables and covariates in the former set do not result from the use of the same testing instrument, whereas in the latter set they do result from the same testing instrument. In this paper, the type of models considered are of the latter type although the estimation techniques discussed here can be adapted to models of the former type (Fielding and Spencer, 1997; Spencer and Fielding, 2002).

The choice of analysing with growth curves or lagged covariates often depends on the aims of the analysis, and also, if the number of occasions for which data are available is small, analysis using growth curves can be difficult. The estimation techniques shown in this paper only need a minimum of 3 or 4 occasions, and can be adapted to cope with just 2 occasions (Fielding and Spencer, 1997; Spencer and Fielding, 2002). In this paper, it is shown that the standard multilevel estimation techniques generally used for models containing a lagged regressor may not produce consistent parameter estimates. Alternative estimation strategies are proposed.

In section 2 of this paper we introduce the model of interest, and the issue of conventional estimation procedures producing inconsistent results for this type of model is discussed in section 3. In section 4 we develop procedures for consistent estimation and demonstrate them on simulated datasets. The procedures are applicable for two different assumptions about the nature of the covariates in the model. Extensions to the class of models eligible for this treatment are considered in section 5 and some summary conclusions are given in section 6.

2. The model of interest

The model of interest is

$$y_{it} = \beta_0 + \delta_{0i} + \beta_1 y_{i(t-1)} + (\beta_2 + \alpha_{2i})x_{it} + e_{it}$$

$$\delta_{0i} \overset{\text{iid}}{\sim} \text{N}(0, \sigma_\delta^2) \qquad \alpha_{2i} \overset{\text{iid}}{\sim} \text{N}(0, \sigma_{\alpha_2}^2) \qquad e_{it} \overset{\text{iid}}{\sim} \text{N}(0, \sigma_e^2) \tag{1}$$

$$cov(\delta_{0i}, \alpha_{2i}) = \sigma_{\delta\alpha_2} \qquad cov(\delta_{0i}, e_{it}) = 0 \qquad cov(\alpha_{2i}, e_{it}) = 0$$

where y_{it} is an observation on subject i at occasion t, the intercept of the model is $\beta_0 + \delta_{0i}$, allowing it to vary randomly between subjects, and β_1 is the coefficient of the lagged regressor, constant over subjects. The $(\beta_2 + \alpha_{2i})$ coefficient of the x_{it} is allowed to vary randomly between subjects and e_{it} is the subject-occasion specific error term. The δ_{0i} are independently, identically distributed as are the α_{2i} and e_{it}.

We can regard this as being a multilevel model because we have two levels of observation. At the lowest level we have time and we have a number of observations nested within our second level: student.

The model is of interest in educational and cognitive research, as it allows researchers to investigate the effect that a factor has on the change in a test mark from, say, year to year. Without the lagged regressor, the model would be investigating the effect that the other factor has on the mark at a particular time, and not on the change. As one of the basic purposes of education is to bring about change, this distinction is vital. In addition, if no allowance is made for previous educational or cognitive attainment, then the factor (possibly relating to the current year's education) faces the burden of explaining all the attainment of the preceding years as well as the increased attainment in the current year.

The model we are considering thus has two types of covariate, the lagged regressor $(y_{i(t-1)})$ and x_{it}. We consider two assumptions: (1) that the x_{it} are independently, identically distributed and (2) $x_{it} = x_{i(t-1)} + \gamma + v_{it}$ where the v_{it} are independently, identically distributed, $E(v_{it}) = 0$, $cov(v_{it}, \delta_{0i}) = 0$, $cov(v_{it}, \alpha_{2i}) = 0$, $cov(v_{it}, e_{it}) = 0$. In both cases, we additionally assume that x_{it} is uncorrelated with e_{it}. If $\gamma = 0$ for assumption 2 then the variable x_{it} has a random walk over time. If $\gamma \neq 0$, x_{it} has a trend over time. These very general assumptions preclude few variables that might be of interest to researchers. The use of more than one covariate is considered in section 5.

Examining the parts of the model more closely, we may have a series of test scores y_{it} on student i at various time points, denoted by t. The covariate x_{it} may be a variable such as exact age at testing occasion t. The intercept (β_0) and slope (β_1) have the usual regression interpretations. The

126

δ_{0i} may be regarded as a measure of innate student ability, relative to the overall average ability. Students who are innately better at the test and test material will have positive values for δ_{0i} and those who are innately worse than average will have negative values. The δ_{0i} may also allow for constant student characteristics not accounted for in the model covariates. The α_{2i} allows the slope associated with the covariate x_{it} to vary between students. Students who innately have more positive slopes for the covariate will have positive values for α_{2i} and those who innately have less positive slopes will have negative values for α_{2i}. The e_{it} is simple random error.

Because δ_{0i} and α_{2i} are not time-dependent, we have (for both assumption 1 and assumption 2) that $cov(\delta_{0i}, x_{it}) = cov(\delta_{0i}, x_{i(t-1)})$ and $cov(\alpha_{2i}, x_{it}) = cov(\alpha_{2i}, x_{i(t-1)})$ for all t and thus $cov(\delta_{0i}, x_{it}) = \sigma_{\delta x}$ and $cov(\alpha_{2i}, x_{it}) = \sigma_{\alpha_2 x}$ for all t. Also, we have $cov(\alpha_{2i}, x_{it} x_{i(t-k)}) = cov(\alpha_{2i}, x_{it} x_{i(t-k-m)})$ for all t, k, m and thus $cov(\alpha_{2i}, x_{it} x_{i(t-k)}) = \sigma_{\alpha_2 x^2}$ for all t, k.

3. Estimation and the problem of inconsistency

Parameter estimation for multilevel models can be carried out using maximum likelihood methods. There are three algorithms that are commonly used to accomplish this estimation. Goldstein (1986) uses Iterative Generalised Least Squares (IGLS) to obtain maximum likelihood estimates for the parameters in the random part of the model (the variances and covariances). These estimates are then used to obtain estimates of parameters from the fixed part of the model (the intercept and coefficients for the regressors) using Least Squares methods. The IGLS algorithm is implemented in the software MLwiN. Longford (1987) uses a Fisher Scoring algorithm that is equivalent to IGLS. The software VARCL is based on Longford's work. An empirical Bayes estimation method that utilises the EM algorithm has been developed by Raudenbush and Bryk (1986). It is implemented in the software HLM.

All these estimation algorithms incorporate Least Squares methods, and it is this fact that leads to the problem of inconsistency for the model described above. IGLS (and effectively the equivalent Fisher Scoring) uses Least Square methods explicitly to obtain estimates of the fixed part coefficients. The empirical Bayes method also explicitly incorporates Least

Squares estimates into its estimation process.

To demonstrate the problem of inconsistency, let us start by looking at Ordinary Least Squares (OLS) theory which operates by assuming independence of regressors and error. Applying $(X^T X)^{-1} X^T$ to both sides of $y = X\beta + u$ and solving for β, we obtain

$$\beta = (X^T X)^{-1} X^T y - (X^T X)^{-1} X^T u$$

The OLS estimator $\hat{\beta} = (X^T X)^{-1} X^T y$ will thus converge to the true value β if the expected value of $(X^T X)^{-1} X^T u$ is equal to zero. That is, the OLS estimator will be a consistent estimator of β if $E((X^T X)^{-1} X^T u) = 0$ or equivalently (as $(X^T X)^{-1}$ cannot be zero), $E(X^T u) = 0$.

As the algorithms that carry out the maximum likelihood estimation rely on Least Squares methods, they too require $E(X^T u) = 0$ to yield consistent parameter estimates.

Thus considering our model of interest (1), for the estimates of β_0, β_1 and β_2 to be consistent when the parameters are estimated, $E(X^T u)$ must be zero where for $y = X\beta + u$, we have:

$$y = \{y_{it}\}, \quad X = \{1 \ , \ y_{i(t-1)} \ , \ x_{it}\},$$
$$\beta = \{\beta_0, \beta_1, \beta_2\}^T$$
$$u = \{\delta_{0i} + \alpha_{2i} x_{it} + e_{it}\}.$$

With $y_{i(t-1)}$ and x_{it} both being correlated with δ_{0i} and α_{2i}, the requirement for X and u to be independent will not be satisfied and the estimate of β will, in general, be inconsistent.

This problem comes from two sources. One source is the correlations that exist between the x_{it} and the δ_{0i}, α_{2i}. The other source is the correlation between the lagged regressor, $y_{i(t-1)}$ and δ_{0i}, α_{2i}.

If the lagged regressor were not present and we assumed that δ_{0i} and α_{2i} were not correlated with the x_{it}, then the expectations would be zero. The least squares estimates would then be consistent, as in the models considered by Goldstein (1995) and Longford (1987). The model we consider here is thus of a more complex form than that considered by these two authors.

128

4. First and second difference instrument methods

Solutions to this problem of inconsistency caused by a lagged regressor under assumption 1 or 2 have been proposed by Kiviet (1995) and Rice et al (1999). Kiviet uses a bias corrected version of the least squares dummy variable estimator and Rice et al use conditioned iterative generalised least squares. The first of these approaches suffers from a problem in that sometimes the bias correction applied actually increases the bias, and neither of them are capable of being extended to cope with cases where a level one random effect is correlated with a regressor. Here, we follow an approach suggested by Anderson and Hsiao (1981) and take differences so that model (1) becomes:

$$y_{it} - y_{i(t-1)}$$
$$= \beta_1(y_{i(t-1)} - y_{i(t-2)}) + (\beta_2 + \alpha_{2i})(x_{it} - x_{i(t-1)}) + e_{it} - e_{i(t-1)} \quad (2)$$

or $y = X\beta + u$ where $y = \{y_{it} - y_{i(t-1)}\}$, $X = \{y_{i(t-1)} - y_{i(t-2)} \,, \, x_{it} - x_{i(t-1)}\}$, $\beta = \{\beta_1, \beta_2\}^T$, and $u = \{\alpha_{2i}(x_{it} - x_{i(t-1)}) + e_{it} - e_{i(t-1)}\}$.

Now it can be shown that $x_{it} - x_{i(t-1)}$ is independent of the error, u, but $y_{it} - y_{i(t-1)}$ is still not independent of u. We thus have a non-zero $E(X^T u)$ and we will still not obtain a consistent estimator of β despite the differencing. Now, however, the inconsistency is only due to the presence of the lagged regressor $y_{i(t-1)} - y_{i(t-2)}$ in the model and its correlation with α_{2i}, $x_{it} - x_{i(t-1)}$ and $e_{it} - e_{i(t-1)}$. By differencing, we have solved the problem of our x_{it} regressor being correlated with the δ_{0i} and α_{2i}. Thus, if a lagged regressor did not exist then we would be obtaining consistent estimates using standard procedures. This was not the case for the undifferenced model (1) in section 2 when we have δ_{0i}, α_{2i} correlated with the x_{it}.

At this stage we turn to instrumental variable estimation. This method of estimation is frequently used in the field of econometrics to overcome problems of endogeneity, such as are presented here, and its use in connection with multilevel models is briefly discussed by Mason (1995). The essentials of the method are that as it is the relationship between the set of regressors, X, and the error u that is causing difficulties because $E(X^T u) \neq 0$, we use another set of variables, Z, which is closely related to the original set of

129

regressors, X, but has the property $E(Z^T u) = 0$. Applying now $(Z^T X)^{-1} Z^T$ to both sides of $y = X\beta + u$, we obtain $\beta = (Z^T X)^{-1} Z^T y - (Z^T X)^{-1} Z^T u$.

The instrumental variable estimator $\beta^* = (Z^T X)^{-1} Z^T y$ will thus converge to the true value β if the expected value of $(Z^T X)^{-1} Z^T u$ is equal to zero. That is, the instrumental variable estimator will be a consistent estimator of β if $E((Z^T X)^{-1} Z^T u) = 0$ or equivalently (as $(Z^T X)^{-1}$ will not be zero if Z is related to X), $E(Z^T u) = 0$. Bowden and Turkington (1984) provide further details.

For our instrument set, we will choose $x_{it} - x_{i(t-1)}$ as an instrument for itself as it does not cause the inconsistency problems, and we are left to choose an instrument for $y_{it} - y_{i(t-1)}$. It is this action of choosing an instrument which is one of the most problematic parts of instrumental variable estimation. It may be hypothesised that a variable which is highly correlated with the problematic regressor is influenced by similar processes as the regressor. If this is the case then the assumption that the variable is uncorrelated with the model error may be hard to sustain. Here, however, we choose an instrument for the lagged regressor that actually comes from the data itself and no extra variables are considered. That is, the repeated measurements data provide "naturally occurring" instruments. We thus obtain an instrument that is correlated with the lagged regressor, and below we show that it is uncorrelated with the model error.

An intuitive choice of instrument for the regressor $y_{it} - y_{i(t-1)}$ will be one that at least avoids a correlation with the $e_{it} - e_{i(t-1)}$ in the model error. Taking a previous lag, we get $Z_1 = \{y_{i(t-2)} - y_{i(t-3)} , x_{it} - x_{i(t-1)}\}$. Another possible candidate is $Z_2 = \{y_{i(t-2)} , x_{it} - x_{i(t-1)}\}$. Examining the first line of $E(Z_1^T u)$, we have

$$
\begin{aligned}
& E\left[(y_{i(t-2)} - y_{i(t-3)})(\alpha_{2i}(x_{it} - x_{i(t-1)}) + e_{it} - e_{i(t-1)})\right] \\
= & \left[E\left(\alpha_{2i}(y_{i(t-2)} - y_{i(t-3)})(x_{it} - x_{i(t-1)})\right)\right] \\
& \quad + \left[E\left((y_{i(t-2)} - y_{i(t-3)})(e_{it} - e_{i(t-1)})\right)\right] \\
= & \left[E\left(\alpha_{2i}(y_{i(t-2)} - y_{i(t-3)})(\gamma + v_{it})\right)\right] \\
& \quad + \left[E\left((y_{i(t-2)} - y_{i(t-3)})(e_{it} - e_{i(t-1)})\right)\right] \\
= & \left[E\left(\alpha_{2i}(y_{i(t-2)} - y_{i(t-3)})\gamma\right)\right] + \left[E\left(\alpha_{2i}(y_{i(t-2)} - y_{i(t-3)})E(v_{it})\right)\right]
\end{aligned}
$$

130

$$+ \left[E \left((y_{i(t-2)} - y_{i(t-3)})(e_{it} - e_{i(t-1)}) \right) \right]$$
$$= \gamma E \left(\alpha_{2i}(y_{i(t-2)} - y_{i(t-3)}) \right)$$

as α_{2i} and $y_{i(t-2)} - y_{i(t-3)}$ are independent of $e_{it} - e_{i(t-1)}$ and of $x_{it} - x_{i(t-1)}$ (and thus v_{it}) under assumptions 1 and 2. Also $E(v_{it}) = 0$ under assumption 2 and $E(e_{it} - e_{i(t-1)}) = 0$. Under assumption 1, $E(x_{it} - x_{i(t-1)}) = 0$, so the whole of the above expectation resolves to zero.

The second line of $E(Z_1^T u)$ is

$$E \left((x_{it} - x_{i(t-1)})(\alpha_{2i}(x_{it} - x_{i(t-1)}) + e_{it} - e_{i(t-1)}) \right)$$
$$= \left[E \left(\alpha_{2i}(x_{it} - x_{i(t-1)})^2 \right) \right] + \left[E \left((x_{it} - x_{i(t-1)})(e_{it} - e_{i(t-1)}) \right) \right]$$
$$= 0$$

as α_{2i} and $e_{it)} - e_{i(t-1)}$ are both independent of $x_{it} - x_{i(t-1)}$ under assumptions 1 and 2.

By a similar line of argument, the first line of $E(Z_2^T u) = \gamma E \left(\alpha_{2i} y_{i(t-2)} \right)$ for assumption 2 and equals zero under assumption 1. The expectation in the second line of $E(Z_2^T u)$ equals zero again as it is identical to the second line of $E(Z_1^T u)$.

Thus, if we additionally assume that $\gamma = 0$, the expectations $E(Z_1^T u)$ and $E(Z_2^T u)$ are zero under either assumption 1 or assumption 2, and the instrumental variable estimators will be consistent estimators of β. What we have effectively done here is to cause the regressors and errors to be independent of each other, and so this first difference instrument method automatically satisfies this assumption made in multilevel modelling.

If we cannot assume that $\gamma = 0$ in assumption 2, we proceed by again differencing the first differenced model (2).

$$y_{it} - 2y_{i(t-1)} + y_{i(t-2)}$$
$$= \beta_1(y_{i(t-1)} - 2y_{i(t-2)} + y_{i(t-3)}) + (\beta_2 + \alpha_{2i})(x_{it} - 2x_{i(t-1)}$$
$$+ x_{i(t-2)}) + e_{it} - 2e_{i(t-1)} + e_{i(t-2)} \tag{3}$$

For normal regression techniques to yield consistent parameter estimates for this model, we require independence of regressors and error. It can be shown that this does not occur.

As before, we use instrumental variable estimation and require $E(Z^T u) = 0$ for consistency. Taking $Z = \{y_{i(t-3)} - 2y_{i(t-4)} + y_{i(t-5)}$, $x_{it} - 2x_{i(t-1)} + x_{i(t-2)}\}$ ($y_{i(t-3)} - 2y_{i(t-4)} + y_{i(t-5)}$ being a second difference instrument), $Z = \{y_{i(t-3)}$, $x_{it} - 2x_{i(t-1)} + x_{i(t-2)}\}$ or $Z = \{y_{i(t-3)} - y_{i(t-4)}$, $x_{it} - 2x_{i(t-1)} + x_{i(t-2)}\}$, we do obtain $E(Z^T u) = 0$ (see Appendix) and thus this second difference instrument method produces consistent estimates for the case when $\gamma \neq 0$.

4.1 Implementation of first difference instrument method

The first difference instrument method was implemented as part of the VARCL package. VARCL is based on the estimation procedure suggested by Longford (1987). Part of its algorithm involves estimating the fixed parameters of the model, given estimates of the random parameters. To implement the first difference instrument method, VARCL's usual method of estimating the fixed parameters is replaced by the procedure for instrumental variable estimation outlined above. That is, we will use instrumental variable methods to estimate the parameters β_1 and β_2 from the first differenced model (2). This model does not contain β_0, so this is then estimated from model (1), using overall means and the instrumental variable estimates of β_1 and β_2:

$$\beta_0^* = (\text{Mean of } y_{it}) - [\beta_1^* \times (\text{Mean of } y_{t-1})] - [\beta_2^* \times (\text{Mean of } x_{it})]$$

This complete set of estimates for the fixed parameters is then used in VARCL to estimate the random parameters of model (1).

4.2 Simulated datasets for the first difference instrument method

To demonstrate the use of the first difference instrument method above, we created two groups of simulated datasets, both based on model (1) with the x_{it} defined as in assumption 1. The parameters that group A of the simulated datasets was based on came from an analysis of the National Foundation for Educational Research's "Streaming Longitudinal Study". See Barker Lunn (1970) for more details. The parameters used were $\beta_0 = 3\cdot35$,

$\beta_1 = 0{\cdot}87$, $\beta_2 = 0{\cdot}06$, $\sigma_\delta^2 = 0{\cdot}13$, $\sigma_{\alpha_2}^2 = 0{\cdot}01$, $\sigma_e^2 = 14{\cdot}31$, and $\sigma_{\delta\alpha_2}^2 = -0{\cdot}02$. For group B of the simulated datasets, the same parameters, but with $\beta_1 = 0{\cdot}2$, $\sigma_\delta^2 = 1{\cdot}13$, $\sigma_{\alpha_2}^2 = 1{\cdot}01$ was used. Each group consisted of 50 datasets, each containing a number of "results" for each of 500 "pupils". An arbitrary start point $y_{i0} = 0$ was defined and y_{i1} was simulated. The reading y_{i2} was then created using y_{i1} as the lagged reading and so on, until a plot of the y_{it} revealed that the effect of the arbitrary start point had worn off. The subsequent simulated y_{it} were used as the readings for the dataset.

Although many simulated readings could be taken for each pupil, here we will focus on an analysis with 3 responses per pupil, y_{it}, $y_{i(t-1)}$ and $y_{i(t-2)}$, with associated lagged regressors $y_{i(t-1)}$, $y_{i(t-2)}$ and $y_{i(t-3)}$ and covariates x_{it}, $x_{i(t-1)}$ and $x_{i(t-2)}$ respectively. As can be seen in subsequent sections, although we are primarily interested in analyses involving these readings, previous readings of the y variable can be utilised to aid the estimation process.

4.3 Simulation examples for the first difference instrument method

The first simulation that we carry out looks at the implementation of the first difference instrument method using group A of the simulated datasets. With three responses, y_{it}, $y_{i(t-1)}$ and $y_{i(t-2)}$, taking first differences leaves us with two equations:

$$
\begin{aligned}
&y_{it} - y_{i(t-1)} \\
&= \beta_1\big(y_{i(t-1)} - y_{i(t-2)}\big) + (\beta_2 + \alpha_{2i})(x_{it} - x_{i(t-1)}) + e_{it} - e_{i(t-1)} \qquad (4) \\
&y_{i(t-1)} - y_{i(t-2)} \\
&= \beta_1\big(y_{i(t-2)} - y_{i(t-3)}\big) + (\beta_2 + \alpha_{2i})(x_{i(t-1)} - x_{i(t-2)}) + e_{i(t-2)} - e_{i(t-2)}
\end{aligned}
$$

The instrument set that is used to implement the first difference instrument method for these equations consists of the first difference instruments $y_{i(t-2)} - y_{i(t-3)}$, $x_{it} - x_{i(t-1)}$ for the first equation and the same with just the previous lag $(y_{i(t-3)} - y_{i(t-4)},\ x_{i(t-1)} - x_{i(t-2)})$ for the second. Thus, in addition to the data whose simulation is described in section 4.2, we use a previous lag of the response variable, $y_{i(t-4)}$.

Table 1 shows the results of using the standard VARCL estimation procedure to estimate the model parameters for the datasets in group A, alongside the results produced by using the first difference instrument method with the instruments described above. We see that the first difference instrument method is evidently struggling to carry out the estimation. The efficiency of the estimates is very poor and they are of little use. This is because some extreme parameter estimates are being obtained for some of the simulations by VARCL. Table 2 shows the results obtained when group B of the simulated datasets is used.

Table 1: 1st Difference Instrument Method: Group A, 1st Difference Instrument

	Target Values	Estimates obtained using standard VARCL[a]	Estimates using instrument method[a]
β_0	3·35	2·71 (0·27)	61·67 (183·62)
β_1	0·87	0·90 (0·01)	−1·93 (8·84)
β_2	0·06	0·05 (0·02)	0·27 (0·79)
σ_e^2	14·31	14·47 (0·60)	853·9 (4070·0)
σ_δ^2	0·13	0·16 (0·17)	4287·8 (18919·9)
$\sigma_{\alpha_2}^2$	0·01	0·01 (0·005)	30·10 (142·14)
$\sigma_{\delta\alpha_2}$	−0·02	−0·01 (0·03)	−31·40 (135·32)

[a]Estimates are means from 50 simulations. The se in brackets are empirical standard deviations for the 50 simulations.

Looking at the question of consistency, both tables 1 and 2 provide evidence that the standard estimation procedure produces inconsistent results, but for the first difference instrument method, the mean parameter estimates now fall well within two standard deviations of the "target" values that the data were simulated from. Thus we have some evidence that consistent results are being obtained with the first difference instrument method, as we would expect, although the evidence from Group A is not too convincing given the large standard errors involved. Group B, with the lower standard errors in table 2, provides rather more evidence of consistency.

Table 2: 1st Difference Instrument Method: Group B, 1st Difference Instrument

	Target Values	Estimates obtained using standard VARCL[a]	Estimates using instrument method[a]
β_0	3·35	1·08 (0·49)	3·20 (0·48)
β_1	0·20	0·74 (0·12)	0·23 (0·11)
β_2	0·06	0·01 (0·03)	0·02 (0·17)
σ_e^2	14·31	22·24 (2·60)	14·75 (1·34)
σ_δ^2	1·13	0·02 (0·03)	1·10 (1·40)
$\sigma_{\alpha_2}^2$	1·01	0·12 (0·23)	1·01 (0·29)
$\sigma_{\delta\alpha_2}$	−0·02	0·0002 (0·05)	0·06 (0·22)

[a]Estimates are means from 50 simulations. The se in brackets are empirical standard deviations for the 50 simulations.

4.4 Regarding the efficiency of instrumental variable methods

Bowden and Turkington (1984, pp. 29-32) show that in instrumental variable estimation, greater efficiency is obtained when the canonical correlations between Z and X are increased. Here, Z and X both always contain $x_{it} - x_{i(t-1)}$, so one of the canonical correlations will always be exactly one. This means that the second (and last, given that the dimension of Z and X is two) canonical correlation will give us an indication of the efficiency we can expect. These canonical correlations have been calculated for each dataset. For group A of the simulated datasets, the mean second canonical correlation over the fifty datasets is 0·03, whereas for group B it is 0·37. Thus, we would expect group B to give the much more efficient results that it does.

Robertson and Symons (1992) deal with the issue of the regressor and instrument having different correlations depending on the true parameter values. These results show that the efficiency of the estimation procedure depends heavily on the size of β_1. Looking at the figures in table 1 which result from using the first difference instrument method with group A of the

135

simulated datasets, we see that although we are obtaining consistent estimates, the relatively large associated standard errors make any meaningful interpretation of the parameter estimates virtually impossible. Indeed, it could be argued that as the biases involved are not excessive, the inconsistent parameter estimates are preferable to the consistent estimates as their precision is substantially better.

However, with a different value for β_1 and, as a result, a higher second canonical correlation between the design matrix and the instrument set, we see that the standard errors in table 2, when group B of the simulated datasets is used, are not excessive, and the results obtained show evidence of the consistency and are interpretable.

This ability to look at the canonical correlations and obtain a measure of the efficiency of parameter estimation is of great practical use. Not only can we decide which instrument to use (see section 4.6), but before any work has been done, we can have some idea of the efficiency of the parameter estimation and thus if it is worth using the methods detailed here.

Instead of using first differenced instruments, we now look at the effects of using undifferenced instruments. For the first differenced model equations (4) in section 4.3, using undifferenced instruments means using $y_{i(t-2)}$ for the first equation and $y_{i(t-3)}$ for the second. We thus obtain the parameter estimates for groups A and B of the simulated datasets shown in table 3.

Once again, the estimates obtained for group A are extremely inefficient, as we would expect given that the mean second canonical correlation here is 0·08, but are slightly better than when the first difference instrument was used in table 1 and the mean second canonical correlation was 0·03. For group B, the mean second canonical correlation is 0·22 and the results are more efficient than for group A, although they are not as efficient as when the first difference instrument was used and the mean second canonical correlation was 0·37. Again, these results for group B provide some evidence of the consistency we expect. The extremely large mean estimate of σ_e^2 is caused by a few "rogue" simulations giving very large estimates.

4.5. Simulated datasets for the second difference instrument method

Table 3: 1st Difference Instrument Method: Groups A/B, Undifferenced Instrument

	Group A of Simulated Datasets		Group B of Simulated Datasets	
	Target Values	Estimates using instrument method[a]	Target Values	Estimates using instrument method[a]
β_0	3·35	$-0·24$ (13·53)	3·35	3·29 (0·56)
β_1	0·87	1·04 (0·65)	0·20	0·21 (0·13)
β_2	0·06	0·02 (0·20)	0·06	0·02 (0·17)
σ_e^2	14·31	32·55 (83·17)	14·31	207965·5 (1446779·3)
σ_δ^2	0·13	42·61 (153·6)	1·31	1·72 (1·97)
$\sigma_{\alpha_2}^2$	0·01	0·21 (0·44)	1·01	1·07 (0·48)
$\sigma_{\delta\alpha_2}$	$-0·02$	$-0·31$ (0·82)	$-0·02$	0·20 (0·61)

[a]Estimates are means from 50 simulations. The se in brackets are empirical standard deviations for the 50 simulations.

In order to demonstrate the success of the second difference instrument method, we create two more groups of simulated datasets. Group C is the same as group A of section 4.2, except that now we define the covariate x_{it} by $x_{it} = x_{i(t-1)} + \gamma + v_{it}$, $v_{it} \sim N(0·0,1·0)$, $\gamma = 1$ with the v_{it} independently, identically distributed. This creates a trend over time in the x_{it}. The parameter values we use to simulate these datasets are the same as we use for group A. Group D of the simulated datasets is created in the same manner as group B of section 4.2, with the addition of a trend over time in the x_{it} as above for group C.

4.6 Simulation examples for the second difference instrument method

With the three responses y_{it}, $y_{i(t-1)}$ and $y_{i(t-2)}$ of our simulated datasets, taking second differences leaves us with just equation (3).

The instrument sets that can be used to implement the second difference instrument method are as described in section 4. All three contain the

second difference in the covariate $(x_{it} - 2x_{i(t-1)} + x_{i(t-2)})$ and involve $y_{i(t-3)}$, but to examine the first differenced and second differenced instruments, we also use the previous lags $y_{i(t-4)}$ and $y_{i(t-5)}$. As described in section 4.4, the decision as to which instrument to use can be made by looking at the canonical correlations between the original regressors in the model and the proposed instrument set. For group C of the simulated datasets, the second canonical correlations are 0·033, 0·039 and 0·036 for the undifferenced, first difference and second difference instruments respectively. For group D of the simulated datasets, the second canonical correlations are 0·098, 0·163 and 0·084 respectively. Thus, as the first difference instrument gives the highest canonical correlation for both groups of simulated datasets, we choose it for use in the analysis.

The second difference instrument method was implemented as part of the VARCL package in the same way as the first difference instrument method.

Estimating the parameters of the simulated datasets in groups C and D with VARCL and using the second difference instrument method with the first difference instrument yields table 4 for group C and table 5 for group D. Also shown are the results of using the standard VARCL estimation method to obtain the parameter estimates.

In both tables 4 and 5, we see evidence that the standard estimation procedure gives inconsistent estimates, whereas the second difference instrument method produces estimates that do not point towards inconsistency. As with the first difference method, we see that where β_1 has a target value of 0·87 (group C), the efficiency of the estimates produced for the instrument method (table 4) is poor. The empirical standard deviations are very large, and due to this, little can be deduced about the consistency of the parameter estimates. In table 5, we see much smaller empirical standard deviations, as we would expect with the higher canonical correlation, and have more evidence of consistency.

5. Extension to more non-lagged regressors and levels

Typically in an analysis, more than one non-lagged regressor would be required in the model. Also, more than two levels in the model hierarchy would be required. This would allow us to take account of, for instance, the

Table 4: 2nd Difference Instrument Method: Group C, 1st Difference Instrument

	Target Values	Estimates obtained using standard VARCL[a]	Estimates using instrument method[a]
β_0	3·35	2·92 (0·29)	87·15 (257·1)
β_1	0·87	0·89 (0·01)	−3·10 (12·26)
β_2	0·06	0·05 (0·02)	0·05 (1·03)
σ_e^2	14·31	14·42 (0·60)	1653·5 (6381·5)
σ_δ^2	0·13	0·30 (0·26)	9473·2 (36703·9)
$\sigma_{\alpha_2}^2$	0·01	0·01 (0·005)	66·83 (249·3)
$\sigma_{\delta\alpha_2}$	−0·02	0·03 (0·03)	−45·92 (179·5)

[a]Estimates are means from 50 simulations. The se in brackets are empirical standard deviations for the 50 simulations.

grouping of pupils into classes and then into schools.

If we allow there to be m levels and p non-lagged regressors, we get:

random effects: $\delta_{0I1}, \delta_{0I2}, \ldots, \delta_{0Im}, e_{It}$

regressors: $y_{I(t-1)}, x_{It1}, \ldots, x_{Itp}$

coefficients: $\beta_1, \beta_{21} + \alpha_{2I11} + \alpha_{2I21} + \ldots + \alpha_{2Il1}, \ldots,$
$\beta_{2p} + \alpha_{2I1p} + \alpha_{2I2p} + \ldots + \alpha_{2Ilp}$

where the suffix I represents a combination of suffixes to indicate the various levels.

139

Table 5: 2nd Difference Instrument Method: Group D, 1st Difference Instrument

	Target Values	Estimates obtained using standard VARCL[a]	Estimates using instrument method[a]
β_0	3·35	0·35 (0·14)	2·83 (2·72)
β_1	0·20	0·93 (0·02)	0·30 (0·37)
β_2	0·06	0·01 (0·02)	0·06 (0·32)
σ_e^2	14·31	26·67 (1·14)	322·3 (1451·1)
σ_δ^2	1·13	0·76 (0·13)	52·60 (45·03)
$\sigma_{\alpha_2}^2$	1·01	0·01 (0·003)	1·43 (1·17)
$\sigma_{\delta\alpha_2}$	−0·02	0·09 (0·02)	7·85 (6·47)

[a]Estimates are means from 50 simulations. The se in brackets are empirical standard deviations for the 50 simulations.

Then:

$$E(X^T u) = E \begin{pmatrix} \left(\sum_{j=1}^{m} \delta_{0Ij} \right) + \left(\sum_{j=1}^{m} \sum_{k=1}^{p} \alpha_{2Ijk} x_{Itk} \right) + e_{It} \\ y_{I(t-1)} \left[\left(\sum_{j=1}^{m} \delta_{0Ij} \right) + \left(\sum_{j=1}^{m} \sum_{k=1}^{p} \alpha_{2Ijk} x_{Itk} \right) + e_{It} \right] \\ x_{It1} \left[\left(\sum_{j=1}^{m} \delta_{0Ij} \right) + \left(\sum_{j=1}^{m} \sum_{k=1}^{p} \alpha_{2Ijk} x_{Itk} \right) + e_{It} \right] \\ \vdots \\ x_{Itp} \left[\left(\sum_{j=1}^{m} \delta_{0Ij} \right) + \left(\sum_{j=1}^{m} \sum_{k=1}^{p} \alpha_{2Ijk} x_{Itk} \right) + e_{It} \right] \end{pmatrix}$$

The relationship between $E(X^T u)$ and $E(Z^T u)$ in this case of arbitrary numbers of non-lagged regressors and levels is of the same form as when we have one non-lagged regressor and two levels.

Each row of the expectation can, as before, be broken down into sums of expectations of combinations of regressors and random parts. Each expec-

140

tation will, as before, be only linear in the random parts, and thus identical methods can be used to evaluate the expectation and obtain the first and second difference instrument methods.

6. Conclusions

In this paper we have looked in detail at the issue of consistency when one of the regressors in a multilevel model is a previous value of the response variable. The reason for inconsistency, and thus biased results, has been identified.

A new estimation procedure for the fixed parameters has been developed and implemented alongside commercially written software. This first difference instrument method has been shown to be successful in obtaining consistent parameter estimates in theory for two different assumptions about the non-lagged regressors, and evidence of this success has been demonstrated with the use of simulated data. For non-lagged regressors containing a trend (our third assumption), a second difference instrument method has been developed. It has been implemented in software and again evidence of its success in obtaining consistent estimates has been demonstrated using simulated data. The three assumptions made cover a wide range of possible regressors.

The methods used here solve both the problems caused by a lagged regressor and those caused by any correlations existing between the random components of the model and the regressors.

The "success" of the first and second difference instrument methods is tempered by the lack of efficiency that sometimes accompanies them. The usefulness of the estimates is dependent on the canonical correlations between the regressors and instruments used. These canonical correlations can be evaluated before the methods are used, and thus the best set of instruments can be chosen. In addition some idea of the efficiency with which the parameters will be estimated can be gained before using the methods. Decisions can then be made as to whether the results that will be obtained are likely to be meaningful.

Extensions to the first and second difference instrument methods to deal with more than one non-lagged regressor and more than two levels in the

model have been discussed.

Acknowledgements

I am grateful to David Boniface, Richard Davies, Antony Fielding, Teresa Prevost, Catherine Spencer and journal referees and editors for comments on earlier versions of this paper.

Appendix

If we take $Z = \{y_{i(t-3)} - 2y_{i(t-4)} + y_{i(t-5)} \ , \ x_{it} - 2x_{i(t-1)} + x_{i(t-2)}\}$, we get:

$$E(Z^T u) = E \left(\begin{array}{c} A \\ B \end{array} \right)$$

where

$$
\begin{aligned}
A = {} & (y_{i(t-3)} - 2y_{i(t-4)} + y_{i(t-5)}) \times \\
& (\alpha_{2i}(x_{it} - 2x_{i(t-1)} + x_{i(t-2)}) + e_{it} - 2e_{i(t-1)} + e_{i(t-2)}) \\
B = {} & (x_{it} - 2x_{i(t-1)} + x_{i(t-2)}) \times \\
& (\alpha_{2i}(x_{it} - 2x_{i(t-1)} + x_{i(t-2)}) + e_{it} - 2e_{i(t-1)} + e_{i(t-2)})
\end{aligned}
$$

The first line of $E(Z^T u)$ is:

$$
\begin{aligned}
& E(A) \\
&= \left[E \left((y_{i(t-3)} - 2y_{i(t-4)} + y_{i(t-5)})(\alpha_{2i}(x_{it} - 2x_{i(t-1)} + x_{i(t-2)})) \right) \right] \\
& \quad + \left[E \left((y_{i(t-3)} - 2y_{i(t-4)} + y_{i(t-5)})(e_{it} - 2e_{i(t-1)} + e_{i(t-2)}) \right) \right] \\
&= \left[E \left((y_{i(t-3)} - 2y_{i(t-4)} + y_{i(t-5)})(\alpha_{2i}(v_{it} - v_{i(t-1)})) \right) \right] \\
& \quad + \left[E \left((y_{i(t-3)} - 2y_{i(t-4)} + y_{i(t-5)})(e_{it} - 2e_{i(t-1)} + e_{i(t-2)}) \right) \right] \\
&= \left[E \left(\alpha_{2i}(y_{i(t-3)} - 2y_{i(t-4)} + y_{i(t-5)}) \right) E \left(v_{it} - v_{i(t-1)} \right) \right] \\
& \quad + \left[E \left(y_{i(t-3)} - 2y_{i(t-4)} + y_{i(t-5)} \right) E \left(e_{it} - 2e_{i(t-1)} + e_{i(t-2)} \right) \right] \\
&= 0
\end{aligned}
$$

as α_{2i} and $y_{i(t-3)} - 2y_{i(t-4)} + y_{i(t-5)}$ are independent of $x_{it} - 2x_{i(t-1)} + x_{i(t-2)}$ (and thus $v_{it} - v_{i(t-1)}$) and of $e_{it} - 2e_{i(t-1)} + e_{i(t-2)}$. Also $E(v_{it} - v_{i(t-1)}) = 0$ under assumption 3 and $E(e_{it} - 2e_{i(t-1)} + e_{i(t-2)}) = 0$.

The second line of $E(Z^T u)$ is:

$$
\begin{aligned}
&E(B) \\
&= E(\alpha_{2i} x_{it}^2 - 4\alpha_{2i} x_{it} x_{i(t-1)} + 2\alpha_{2i} x_{it} x_{i(t-2)}^2 + 4\alpha_{2i} x_{i(t-1)}^2 \\
&\quad\; - 4\alpha_{2i} x_{i(t-1)} x_{i(t-2)} + \alpha_{2i} x_{i(t-2)}^2 + x_{it} e_{it} - 2x_{it} e_{i(t-1)} + x_{it} e_{i(t-2)} \\
&\quad\; - 2x_{i(t-1)} e_{it} + 4x_{i(t-1)} e_{i(t-1)} - 2x_{i(t-1)} e_{i(t-2)} \\
&\quad\; + x_{i(t-2)} e_{it} - 2x_{i(t-2)} e_{i(t-1)} + x_{i(t-2)} e_{i(t-2)}) \\
&= \sigma_{\alpha_2 x^2} - 4\sigma_{\alpha_2 x^2} + 2\sigma_{\alpha_2 x^2} + 4\sigma_{\alpha_2 x^2} - 4\sigma_{\alpha_2 x^2} + \sigma_{\alpha_2 x^2} \\
&\quad\; + 0 - 0 + 0 - 0 + 0 - 0 + 0 - 0 + 0 \\
&= 0
\end{aligned}
$$

given that $cov(\alpha_{2i}, x_{it} x_{it'}) = \sigma_{\alpha_2 x^2}$, $cov(x_{it}, e_{it'}) = 0$ for all t, t'.

Thus, using this form of Z with the second differenced model, we should obtain consistent estimates. This comes about because now, the $v_{it} - v_{i(t-1)}$ are not correlated with the y terms and can be separated out of the expectation, and also our choice of instrument means what correlations with the e no longer occur. When we take $Z = \{y_{i(t-3)} \;,\; x_{it} - 2x_{i(t-1)} + x_{i(t-2)}\}$, we get:

$$
\begin{aligned}
&E(Z^T u) \\
&= E \begin{pmatrix} y_{i(t-3)}(\alpha_{2i}(x_{it} - 2x_{i(t-1)} + x_{i(t-2)}) + e_{it} - 2e_{i(t-1)} + e_{i(t-2)}) \\ B \end{pmatrix} \\
&= \begin{pmatrix} E\left(y_{i(t-3)}(\alpha_{2i}(x_{it} - 2x_{i(t-1)} + x_{i(t-2)}) + e_{it} - 2e_{i(t-1)} + e_{i(t-2)})\right) \\ 0 \end{pmatrix}
\end{aligned}
$$

given the above result for the second line.

The expectation left is now:

$$
\begin{aligned}
&E\left(y_{i(t-3)}(\alpha_{2i}(x_{it} - 2x_{i(t-1)} + x_{i(t-2)}) + e_{it} - 2e_{i(t-1)} + e_{i(t-2)})\right) \\
&= \left[E\left(y_{i(t-3)}(\alpha_{2i}(x_{it} - 2x_{i(t-1)} + x_{i(t-2)}))\right)\right] \\
&\quad\; + \left[E\left(y_{i(t-3)}(e_{it} - 2e_{i(t-1)} + e_{i(t-2)})\right)\right] \\
&= \left[E\left(y_{i(t-3)}(\alpha_{2i}(v_{it} - v_{i(t-1)}))\right)\right] \\
&\quad\; + \left[E\left(y_{i(t-3)}(e_{it} - 2e_{i(t-1)} + e_{i(t-2)})\right)\right] \\
&= \left[E\left(\alpha_{2i} y_{i(t-3)}\right) E\left(v_{it} - v_{i(t-1)}\right)\right]
\end{aligned}
$$

143

$$+ \left[E\left(y_{i(t-3)}\right) E\left(e_{it} - 2e_{i(t-1)} + e_{i(t-2)}\right) \right]$$
$$= 0$$

as α_{2i} and $y_{i(t-3)}$ are independent of $v_{it} - v_{i(t-1)}$ and of $e_{it} - 2e_{i(t-1)} + e_{i(t-2)}$. Also $E(v_{it} - v_{i(t-1)}) = 0$ under assumption 3 and $E(e_{it} - 2e_{i(t-1)} + e_{i(t-2)}) = 0$.

Again, the consistency occurs because our instrument is sufficiently lagged not to be correlated with the $v_{it} - v_{i(t-1)}$ or the e.

When $Z = \{y_{i(t-3)} - y_{i(t-4)} , x_{it} - 2x_{i(t-1)} + x_{i(t-2)}\}$:

$$E(Z^T u) = E \begin{pmatrix} A_1 \\ B \end{pmatrix} = \begin{pmatrix} E(A_1) \\ 0 \end{pmatrix}$$

where

$$A_1$$
$$= (y_{i(t-3)} - y_{i(t-4)})(\alpha_{2i}(x_{it} - 2x_{i(t-1)} + x_{i(t-2)}) + e_{it} - 2e_{i(t-1)} + e_{i(t-2)})$$

given the above result for the second line.

The expectation left is now:

$$E(A_1)$$
$$= \left[E\left((y_{i(t-3)} - y_{i(t-4)})(\alpha_{2i}(x_{it} - 2x_{i(t-1)} + x_{i(t-2)}))\right) \right]$$
$$\quad + \left[E\left((y_{i(t-3)} - y_{i(t-4)})(e_{it} - 2e_{i(t-1)} + e_{i(t-2)})\right) \right]$$
$$= \left[E\left((y_{i(t-3)} - y_{i(t-4)})(\alpha_{2i}(v_{it} - v_{i(t-1)}))\right) \right]$$
$$\quad + \left[E\left((y_{i(t-3)} - y_{i(t-4)})(e_{it} - 2e_{i(t-1)} + e_{i(t-2)})\right) \right]$$
$$= \left[E\left(\alpha_{2i}(y_{i(t-3)} - y_{i(t-4)})\right) E\left(v_{it} - v_{i(t-1)}\right) \right]$$
$$\quad + \left[E\left(y_{i(t-3)} - y_{i(t-4)}\right) E\left(e_{it} - 2e_{i(t-1)} + e_{i(t-2)}\right) \right]$$
$$= 0$$

as α_{2i} and $y_{i(t-3)} - y_{i(t-4)}$ are independent of $v_{it} - v_{i(t-1)}$ and of $e_{it} - 2e_{i(t-1)} + e_{i(t-2)}$. Also $E(v_{it} - v_{i(t-1)}) = 0$ under assumption 3 and $E(e_{it} - 2e_{i(t-1)} + e_{i(t-2)}) = 0$.

Again we expect to obtain consistent results due to our choosing a suitable instrument which avoids correlations with $v_{it} - v_{i(t-1)}$ and the e.

References

Aitkin, M., Anderson, D. and Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society, Series A* **144**, 419-461.

Aitkin, M. and Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society, Series A* **149**, 1-43.

Anderson, T.W. & Hsiao, C. (1981). Estimation of dynamic models with error components. *Journal of the American Statistical Association* **76**, 375, 598-606.

Barker Lunn, J. C. (1970). *Streaming in the Primary School.* National Foundation for Educational Research in England and Wales: Slough, England.

Bowden, R.J. & Turkington, D.A. (1984). *Instrumental Variables.* Cambridge University Press: Cambridge.

Bryk, A.S. & Raudenbush, S.W. (1992). *Hierarchical linear models.* Sage: Newbury Park, California.

Chock, P. D., Terrell, T. R. & Levitt, S. B. (1975). Time series analysis of Riverside, California, air quality data. *Atmospheric Environment* **9**, 978-989.

Fielding, A. & Spencer, N. H. (1997). Modelling cost-effectiveness in general certificate of education A-level: Some practical methodological innovations. In *Good statistical practice: proceedings of the 12th international workshop on statistical modelling, Biel/Bienne, July 7 to 11, 1997* (eds. C. E. Minder and H. Friedl). Schriftenreihe der Österreichischen Statistischen Gesellschaft: Vienna.

Gilthorpe, M.S., Maddick, I.H. & Petrie, A. (2000). Introduction to multilevel modelling in dental research. *Community Dental Health* **17**, 222-226.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika* **73**, 43-56.

Goldstein, H. (1995). *Multilevel Statistical Models.* Edward Arnold: London.

Goldstein, H. & Thomas, S. (1996). Using examination results as indicators of school and college performance. *Journal of the Royal Statistical Society, Series A* **150**, 149-163.

Hox, J.J. (2002). *Multilevel Analysis: techniques and applications.* Lawrence Erlbaum Associates, Inc.: Mahwah, New Jersey.

Kiviet, J. F. (1995). On bias, inconsistency, and efficiency of various estimators in dynamic panel data models. *Journal of Econometrics* **68**, 53-78.

Kreft, I.G.G. & de Leeuw, J. (1998). *Introducing Multilevel Modeling.* Sage: Newbury Park, California.

Longford, N.T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika* **74**, 4, 817-827.

Longford, N.T. (1993).*Random Coefficient Models.* Clarendon Press: Oxford.

Mason, W. M. (1995). Comment. *Journal of Educational and Behavioral Statistics* **20**, 221-227.

Mason, W. M., Wong, G. Y. and Entwisle, B. (1984). The multilevel linear model: A better way to do contextual analysis. In *Sociological methodology* (ed. K. F. Schuessler), pp. 72-103. Jossey Bass: London.

Milionis, A. E. & Davies, T. D. (1994). Regression and stochastic models for air pollution - I. review, comments and suggestions. *Atmospheric Environment* **28**, 2801-2810.

Paterson, L. (1991). An introduction to multilevel modelling. In *Schools, classrooms and pupils: international studies of schooling from a multilevel perspective* (eds. S. W. Raudenbush and J. D. Willms). Academic Press: San Diego.

Paterson, L. & Goldstein, H. (1991). New statistical methods for analysing social structures: an introduction to multilevel models. *British Educational Research Journal* **17**, 4, 387-393.

Plewis, I. (1991). Using multilevel models to link educational progress with curriculum coverage. In *Schools, classrooms and pupils: international studies*

of schooling from a multilevel perspective (eds. S. W. Raudenbush and J. D. Willms). Academic Press: San Diego.

Plewis, I. (1996). Statistical methods for understanding cognitive growth: a review, a synthesis and an application. *British Journal of Mathematical and Statistical Psychology* **49**, 25-42.

Raudenbush, S. W. and Bryk, B. S. (1986). A hierarchical model for studying school effects. *Sociology of Education* **59**, 1-17.

Rice, N., Jones, A. & Goldstein, H. (1999). Multilevel models where the random effects are correlated with the fixed predictors: a conditioned iterative generalised least squares estimator (CIGLS). University of York, Centre for Health Economics, York University: York.

Robertson, D. and Symons, J. (1992). Some strange properties of panel data estimators. *Journal of Applied Econometrics* **7**, 175-189.

Rutter, C. M. & Elashoff, R. M. (1994). Analysis of longitudinal data: random coefficient regression modelling. *Statistics in Medicine* **13**, 1211-1231.

Snijders, T.A.B. & Bosker, R. (1999). *Multilevel Analysis: an introduction to basic and advanced multilevel modeling.* Sage: Thousand Oaks, California.

Spencer, N. H. & Fielding, A. (2002). A comparison of modelling strategies for value-added analyses of educational data. *Computational Statistics* **17**, 1, 103-116.

Williamson, G. L., Appelbaum, M. and Epanchin, A. (1991). Longitudinal analyses of academic achievement. *Journal of Educational Measurement* **28**, 61-76.

Neil H. Spencer
Department of Statistics, Economics, Accounting and Management Systems
University of Hertfordshire, Hertford Campus
Mangrove Road, Hertford, SG13 8QF, U.K.
N.H.Spencer@herts.ac.uk