THE UNIVERSITY OF HERTFORDSHIRE

BUSINESS SCHOOL

WORKING PAPER SERIES

University of Hertfordshire Business School
College Lane
Hatfield
Hertfordshire
AL10 9AB
United Kingdom

Estimation and Comparison of Endogenous Ordered Category Multilevel Models

Neil Spencer
Business School, University of Hertfordshire,

Antony Fielding
Department of Economics, University of Birmingham,

Abstract

Data often take the form of ordered categories. For instance, in education, test results are often reported as grades. Where a hierarchical structure exists for the data, multilevel modelling of such ordered categorisations can be carried out using macros in MLwiN. The ordered categorisation can be seen as the realisation of an unknown underlying latent variable. A link function is used to relate the two and this determines the scale of the latent variable. This causes a difficulty because whatever model is fitted to the data, the latent variable is rescaled to have the same variance, meaning that developments in the parameter estimates for different models cannot be followed. A heuristic way of overcoming this difficulty has been used by Fielding (2003), using Conditional Mean Scoring (CMS), for models where regressors are not related to the random part of the multilevel model (the exogenous case). In this paper the endogenous case is examined. The use of an instrumental variable approach to overcoming the estimation problems associated with endogenous variables together with the CMS method is shown to be successful in producing a method that allows successive models to be compared. Simulated data and a practical application are used.

1. Introduction

Data collected from many structures often take the form of ordered categories. This is particularly true in education where test results are often in the form of grades that form an ordered categorisation of an underlying continuous scale (e.g. a percentage converted into a grading system). Fielding (1999) discusses the use of ordered categories in modelling educational processes.

In this setting the structure from which the data is collected is organised in a hierarchical manner with pupils grouped into classes which are grouped within schools. Multilevel modelling is an appropriate technique to use to analyse such hierarchically structured data, and the analysis of ordered categorisations can be carried out using macros in the package MLwiN (Yang et al, 2001). Fielding (2002) and Fielding & Yang (2004) discuss such models in detail.

More generally, it can be useful to view an ordered categorisation as the realisation of an underlying latent variable. In the case of an assessment given a mark (e.g. percentage) which is then converted into a grade, or indeed in the case of an assessment which is directly given a grade, the underlying latent variable can be thought of as the student's ability at whatever the assessment is measuring.

The ability of the student is brought about by a combination of influences, and the make-up of the latent variable will mirror this. One of these influences is the teaching that the student has experienced; another is the influence that the school's atmosphere and policies have on the student. Home circumstances and experiences may also have an influence on the grade a student receives, and student characteristics and innate ability (independent of outside influences) will also have an effect.

In the not uncommon situation where the student has had just one teacher involved in the learning associated with the data collected, the effects of teacher and school cannot be completely separated. This can be overcome to some extent where data on teacher and school characteristics exist and can be included in the modelling of the latent variable.

Similarly, home circumstances and experiences, student characteristics and innate ability cannot be completely separated. Again, this can be overcome to some extent by including data on home, family and student in the modelling.

As the latent variable cannot be directly observed, its location, scale and more generally, distribution are arbitrary. For the purposes of analysis, a common choice for the distribution of the latent variable is the standard Normal distribution which gives a location of zero and scale of one. The probit link function connects the latent variable to the observed ordered categories through cumulative probabilities of observing a value in a category at or below a certain point. Adopting the same notation as Fielding (1999), the probability that student *i* in school *j* obtains grade *s* is $\pi_{ij}^{(s)}$. The cumulative probability that this student obtains grade s or less can then be defined as $\gamma_{ij}^{(s)} = \sum_{h=1}^{s} \pi_{ij}^{(h)}$.

These cumulative probabilities can be modelled with the use of a link function: a random effects model would appear as $F^{-1}\left(\gamma_{ij}^{(s)}\right) = \theta_s + (\boldsymbol{X}\beta)_{ij} + u_{0j}$ where F^{-1} is the link function, $\boldsymbol{X}$ is a matrix of regressors, u_{0j} is the random effect of the school and the θ_s (*s* = 1, 2, ..., k–1 where there are k categories) are thought of as cut-points on the latent variable scale (with $\theta_1 < \theta_2 < \mathrm{K} < \theta_{k-1}$).

The matrix of regressors, ***X***, may include data on school, teacher, student, and home and family circumstances and also possibly measures of prior ability. The random effect of the school, u_{0j}, exists to allow for influences not included in the regressors and also includes teacher/school effects.

Another commonly used distribution is the standard logistic distribution, used with the logit link.

2. Building a Model

The starting point for a model building exercise may well be the fitting of a null model – in this case just parameters for the cut-points corresponding to the boundaries of the ordered categories. This null model will have a variance for the latent variable defined by the link function. That is 1 for the probit link, $\pi^2/3$ for the logit link.

The next stage in the model building may be to add a set of explanatory variables, and in conventional modelling this would cause a reduction in the residual variance. The new model could then be compared with the null model to assess the improvement in fit of the model to the data.

However, in the model with the explanatory variables, the latent variable is again scaled to have the variance defined by the link function. In order for this to happen, the cut-points are rescaled and direct comparison with the null model is made impossible. This problem is discussed by Snijders & Bosker (1999).

3. Conditional Mean Scoring and Instrumental Variable Estimation

To overcome this rescaling so that comparisons between models can be made, Fielding (2003) has used Conditional Mean Scoring (CMS), for models where regressors are not related to the random part of the multilevel model (the exogenous case).

In this paper, we consider the case where a regressor is used that is related to the random part of the model (the endogenous case). In a educational setting, this may be the case when a previous test mark is used to model a current test mark, and is related to the student-level random part of the model and possibly also the institution-level random part of the model. As well as having the scaling issue to deal with in model building, the endogenous regressor causes additional estimation problems concerning the consistency of the estimated model parameters. Instrumental variable (IV) methods to deal with this additional problem have been shown to be successful (Spencer & Fielding, 2000; Spencer, 2002; Spencer & Fielding, 2002). The basic IV method is outlined in appendix 1, and an application is shown in more detail in section 6. Here we combine the IV and CMS methods to provide consistent estimation and enable model comparisons.

In section 4, we demonstrate the use of the CMS and IV estimation methods for endogenous ordered category multilevel models using simulated data. In section 5, data from schools in Birmingham are introduced, and in section 6, the CMS and IV estimation methods are applied to these data.

4. CMS and IV Estimation with Simulated Data

Fifty datasets were simulated, each consisting of 36 groups of pupils, each group (or school) containing a number of pupils varying between 11 and 33. A latent variable was created for each pupil in each school, made up of the sum of two components: (i) a component relating to the individual pupil (which may be likened to the pupil's innate ability); (ii) a component relating to the school that the pupil is in (which may be likened to the combined effect of teacher, class dynamics and school). Each of these components was generated using a Normal distribution with a mean of zero and a variance of one.

A variable, which may be likened to a prior test score, was also created. This was made up of the sum of (i) a constant; (ii) the latent variable created above (scaled by 0.5); (iii) a random disturbance taken from a N(0,1) distribution. Additionally, a variable correlated to this prior test score but independent of the latent variable was created as the sum of the constant and random disturbance from (i) and (iii). This variable is used as the instrument for the prior test score in the IV estimation procedure.

A "current" test score was then created using the model

$$
\begin{aligned}
(\text{test score})_{ij} = {} & \text{coefficient} \times (\text{centred prior test score})_{ij} \\
& + (\text{school component of latent variable})_{j} \\
& + (\text{pupil component of latent variable})_{ij} \\
& + (\text{random disturbance})_{ij}
\end{aligned}
$$

where the subscripts *i* and *j* relate to pupil *i* in school *j*. The "centred prior test score" is just the original prior test score with its mean subtracted. This means that the test score will itself be centred around zero.

This model (appropriately) has the school effect and pupil effect in it which have also been used to create the prior test score. This means that the prior test score is endogenous in this model.

This "test score" is then divided into 11 categories akin to grades by defining the cut-points between grades. The 10 cut-points used are shown in table 1. They are evenly spaced between –2.150 and +2.150. With the "test score" having a symmetric (Normal) distribution, this will mean that the distribution of grades will follow a unimodal symmetric distribution. The observed test results are assumed to be these grades in the subsequent analyses.

Having created the fifty datasets, results of fitting the above model to each of them were obtained using the multilevel modelling package MLwiN and an adaptation of the MULTICAT macros (Yang et al, 2001) that enables the probit link function to be used. This adapted version of the macros can be obtained from the authors (N.H.Spencer@herts.ac.uk, A.Fielding@bham.ac.uk). The estimation procedure used amounts to quasi-likelihood (second-order penalized quasi-likelihood: PQL2).

Table 1 shows the mean obtained for each parameter (together with its standard error in brackets) for (i) no application of CMS or IV; (ii) application of CMS but not IV; (iii) application of IV but not CMS; (iv) application of CMS and IV. It should be noted that although the coefficient for centred prior test has a positive sign, estimates of it will correctly have a negative sign. This is because larger prior test scores will lead to larger current test scores and thus an increased probability of being in a higher category. This means that the cumulative probability of being in a particular ordered category or below will decrease. It is this cumulative probability which is being modelled, hence the negative estimates for the prior test parameter. More generally, any non-cut-point regressor in a model of this type will have the sign of its estimate behave in a similar manner.

Table 1: Mean and standard errors of parameter estimates from fifty simulated datasets

	Values	Method 1	Method 2	Method 3	Method 4
	Used in	CMS not used	CMS used	CMS not used	CMS used
Coefficient	simulations	IV not used	IV not used	IV used	IV used
Cut-point 1	–2.150	–2.477(0.224)	–2.142(0.177)	–1.095(0.120)	–2.124(0.174)
Cut-point 2	–1.672	–1.920(0.222)	–1.660(0.175)	–0.858(0.112)	–1.663(0.169)
Cut-point 3	–1.194	–1.376(0.216)	–1.189(0.175)	–0.620(0.107)	–1.201(0.173)
Cut-point 4	–0.717	–0.831(0.207)	–0.717(0.173)	–0.380(0.105)	–0.733(0.184)
Cut-point 5	–0.239	–0.284(0.208)	–0.245(0.179)	–0.135(0.103)	–0.258(0.194)
Cut-point 6	0.239	0.268(0.201)	0.233(0.177)	0.113(0.098)	0.224(0.192)
Cut-point 7	0.717	0.824(0.197)	0.714(0.176)	0.363(0.091)	0.706(0.181)
Cut-point 8	1.194	1.378(0.202)	1.376(1.264)	0.611(0.095)	1.188(0.183)
Cut-point 9	1.672	1.949(0.199)	2.050(2.595)	0.863(0.098)	1.679(0.179)
Cut-point 10	2.150	2.503(0.190)	2.167(0.171)	1.104(0.099)	2.146(0.174)
Centred prior test	0.800	–1.409(0.056)	–1.219(0.043)	–0.409(0.036)	–0.795(0.053)
School variance	1.000	0.796(0.190)	0.598(0.149)	0.348(0.071)	1.359(0.452)

For method 1 where neither CMS or IV estimation is used, it is not surprising to find that the parameter values used to create the simulated data are not recovered. Method 2 addresses the scaling problem by using CMS, but again, the parameter values are not recovered, demonstrating the effect of ignoring the nature of the endogenous variable. Method 3 uses IV estimation to overcome the endogeneity problem, but now the effect of ignoring the scaling issue can be seen as the parameter values are again not recovered. Finally, method 4, using both CMS and IV estimation does recover the parameter values, demonstrating the success of using the combination of methods.

It should be noted that the standard errors of the estimates obtained are respectable. A common objection to using IV methods is that they yield parameter estimates with large standard errors. This is not the case here.

The application of the CMS method here differs from that used in Fielding (2003). Fielding's approach when analysing a set of data is as follows:

1. Obtain estimates of the parameters of the null model (using just the cut-points as regressors).
2. Baseline conditional mean scores and a baseline variance are obtained using the results from this null model.
3. A second model is fitted.
4. Conditional mean scores and hence a variance are obtained from this second model.
5. A scaling factor is calculated as the square root of the ratio of the baseline variance to the variance from the second model.
6. The parameter values of the second model are scaled by the scaling factor.

Fielding (2003) then uses these parameter values as being comparable with those from the null model. Here, the procedure is the same for steps 1 to 6 and then the following takes place:

7. More conditional mean scores and hence a variance are obtained from the scaled parameters of the second model.
8. A new scaling factor is calculated as the square root of the ratio of the baseline variance to the variance from the scaled second model.
9. The scaled parameter values of the second model are further scaled by the new scaling factor.

Steps 7 to 9 are repeated until a stable set of scaled parameter values is obtained. The stable parameter values are used as being comparable with those from the null model.

So that the results of using methods 2 and 4 can be compared with the parameter values used to create the simulated data, the known parameter values for the null model have been used rather than estimates.

5. Data from Birmingham Schools

The data used in this section comes from 4444 children aged around 7 years in 114 schools in Birmingham. The data collected are gender (GENDER), date of birth, ethnic background, first language, eligibility for free school meals (FSM – dummy variable with 1 indicating eligibility) and whether they had attended at least one full term of nursery education. The 10 ethnic background categories and 12 first language categories contain a large degree of overlap, as one would expect, and their use in the analysis may lead to confounding. To overcome this, experimental analysis was undertaken and 14 compound categories from these two variables were identified (AMCLANG1 – AMCLANG14). The date of birth information was converted into age in months which was then centred on age 84 months (CTRDAGE).

Two school context variables were also created: the percentage of pupils eligible for free school meals (PCTFSM) and the average percentage of baseline assessments that were graded above 2 (AVPCTBASEGT2).

In addition, measures of academic achievement and ability were obtained. Baseline assessments of ability carried out by teachers at the beginning of the school year in four areas of mathematics (number, algebra, shape and space, handling data) and three areas of language and literacy (speaking and listening, reading, writing) with pupils being given a grade of (in descending order) 3, 2, 1, 0 in each of the seven areas . Towards the end of the school year, the pupils took the Key Stage 1 Mathematics Standard Assessment Task. Pupils were given grades from this of (in descending order) 3, 2a, 2b, 2c, 1, 0. We use this variable, having six ordered categories, as the response variable in the modelling.

One school was excluded from the analysis because all its baseline assessments were graded as 1. Because of this, it is hard to believe that the baseline data from this school is accurate. Additionally, 11 pupils were excluded because they were given a value for the a, b or c fine grading of grade 2 when in fact they did not have grade 2. This left 4421 pupils in the dataset.

A fuller analysis of these data is contained in Fielding (1999) from which these details of the dataset are taken.

6. CMS and IV Estimation with Birmingham Data

The first step of the analysis was the fitting of a null model (model A) with 5 cut-points relating to boundaries between the 6 categories of the Key Stage 1 test. From the results of this model, baseline conditional mean scores were calculated and hence the baseline variance that is used in the calculation of the scaling factor for all future models.

Following the example of Fielding (2003), the fitting of four further models is planned: model B will use just the baseline assessment variables in addition to the cut-points; model C will use just the pupil characteristic variables in addition to the cut-points; model D will use both the baseline assessment variables and pupil characteristic variables in addition to the cut-points; model E will use the variables to be used in model D plus the school context variables.

However, before the fitting of these models takes place, attention must be paid to how the IV estimation is to be carried out. This requires that variables are found which are related to the endogenous baseline assessment variables but that at the same time are not associated with the random part of the multilevel model: here the school and pupil effects. From Fielding (1999) and experimentation, it is apparent that there are some variables that have been collected that are not significantly related to the Key Stage 1 test result. These are whether or not a pupil attended at least one full term of nursery school and 10 of the 14 dummy variables associated with the compounding of the ethnic background and first language variables. A multilevel model for each of the baseline assessment variables can thus be created, using these 11 variables as regressors:

(baseline assessment)$_{ij}$ =
coefficient × (nursery schooling indicator)$_{ij}$
+ coefficients × (dummies for 10 compound ethnic background
and first language variables)$_{ij}$
+ (school effect)$_{j}$ + (pupil effect)$_{ij}$

where the subscripts *i* and *j* relate to pupil *i* in school *j*.

A variable that is related to the baseline assessment but unrelated to the random effects due to school and pupil can then be obtained as predictions for the baseline assessment from the model based solely on the fixed part of the model (the coefficients and regressors). This can then be used as an instrument for the baseline assessment in the IV estimation process.

However, a complication arises for this dataset because there are seven endogenous baseline assessments. The efficiency of the estimates produced by the IV estimation is subject to the sizes of the canonical correlations between the set of endogenous variables and the set of instrumental variables (see Spencer (2003) for more discussion of this matter). Each baseline assessment variable will not be perfectly correlated with its instrument variable and the lack of perfect correlation, amassed over seven baseline assessment variables means that the canonical correlations and thus efficiency of estimates will be low.

To overcome this, a principal components analysis of the seven baseline assessments has been carried out and the first component extracted. This first component accounts for 59.9% of the variation in the baseline assessment variables. Just the one component is used in the analysis as the second component can only contribute an additional 8.4% of the variation. An instrumental variable for this first principal component is created using the method outlined above for an individual baseline assessment variable. For further information on IV estimation, see e.g. Bowden & Turkington (1984).

Having resolved issues surrounding the IV estimation process, the fitting of models B, C, D and E can be accomplished using CMS for all four models and additionally IV estimation for models B, D and E (the baseline assessment variables not being included in model C means that IV estimation is not necessary for this model). Results from fitting these models are shown in table 2 with standard errors in brackets. In table 2, AMCLANG2 corresponds to an Afro-Caribbean ethnic background with first language English; AMCLANG11 corresponds to a Chinese ethnic background with first language not English; AMCLANG12 corresponds to a Vietnamese ethnic background with first language not English. All three of these are relative to a White ethnic background with first language English.

It should be noted that, as with the results of the simulations, the standard errors of the estimates obtained are respectable for all models including B, D and E where IV estimation takes place.

Table 2: Parameter estimates and standard errors with use of CMS and IV estimation

Coefficient	Model B	Model C	Model D	Model E
Cut-point 1	–2.013(0.051)	–2.225(0.061)	–2.144(0.067)	–2.306(0.089)
Cut-point 2	–0.848(0.051)	–1.102(0.052)	–0.998(0.067)	–1.218(0.089)
Cut-point 3	–0.202(0.051)	–0.473(0.050)	–0.361(0.067)	–0.616(0.089)
Cut-point 4	0.331(0.051)	0.044(0.050)	0.164(0.067)	–0.120(0.089)
Cut-point 5	1.056(0.051)	0.742(0.051)	0.880(0.067)	0.556(0.089)
1st PC for baseline tests	0.247(0.058)		0.229(0.056)	0.185(0.050)
GENDER		0.007(0.030)	–0.085(0.024)	–0.070(0.021)
FSM		0.309(0.033)	0.291(0.047)	0.193(0.021)
CTRDAGE		–0.061(0.004)	–0.033(0.008)	–0.036(0.007)
AMCLANG2		0.053(0.062)	0.138(0.070)	0.103(0.058)
AMCLANG11		–0.629(0.336)	–0.707(0.142)	–0.639(0.118)
AMCLANG12		–0.765(0.436)	–1.201(0.208)	–1.143(0.171)
PCTFSM				0.006(0.002)
AVPCTBASEGT2				0.030(0.018)
School variance	0.236(0.037)	0.176(0.027)	0.201(0.031)	0.141(0.023)

7. Discussion

The analysis of the simulated data in section 4 demonstrates that both CMS and IV estimation are necessary for model comparisons to be made, and that they can both be applied successfully to endogenous ordered category multilevel models. In section 6, CMS and IV estimation have been successfully applied to a dataset from education. The usefulness of CMS to facilitate model comparisons cannot be denied, but to show the effect of using IV, table 3 gives the results of models fitting models B, D and E without using IV estimation (model C has no need for IV estimation and is thus the same as in table 2).

Table 3: Parameter estimates and standard errors with use of CMS but not IV estimation

Coefficient	Model B	Model C	Model D	Model E
Cut-point 1	–1.948(0.066)	–2.225(0.061)	–1.991(0.068)	–2.257(0.107)
Cut-point 2	–0.794(0.057)	–1.102(0.052)	–0.845(0.060)	–1.165(0.102)
Cut-point 3	–0.134(0.056)	–0.473(0.050)	–0.188(0.059)	–0.539(0.101)
Cut-point 4	0.402(0.056)	0.044(0.050)	0.348(0.059)	–0.030(0.101)
Cut-point 5	1.119(0.058)	0.742(0.051)	1.066(0.060)	0.654(0.101)
1st PC for baseline tests	0.370(0.009)		0.360(0.010)	0.344(0.009)
GENDER		0.007(0.030)	–0.125(0.026)	–0.120(0.025)
FSM		0.309(0.033)	0.141(0.029)	0.126(0.028)
CTRDAGE		–0.061(0.004)	–0.017(0.004)	–0.016(0.004)
AMCLANG2		0.053(0.062)	0.233(0.055)	0.219(0.052)
AMCLANG11		–0.629(0.336)	–0.808(0.297)	–0.765(0.282)
AMCLANG12		–0.765(0.436)	–0.847(0.403)	–0.816(0.383)
PCTFSM				0.006(0.002)
AVPCTBASEGT2				0.070(0.015)
School variance	0.317(0.045)	0.176(0.027)	0.298(0.043)	0.217(0.032)

One of the main effects of not using IV estimation is the increased size of the coefficient for the baseline tests. The impact on this coefficient of adding the pupil characteristic variables (going from model B to model D) is similar in both table 2 and table 3, but the impact on the pupil characteristic variables of the introduction of the baseline tests (going from model C to model D) is much more marked. This is particularly true for FSM – with IV estimation, the coefficient only reduces slightly, but without IV estimation, the coefficient is more than halved. Whether IV estimation is used or not also has an effect when adding the school context variables (going from model D to model E). With IV estimation, the coefficient for AVPCTBASEGT2 fails to be significantly different from zero when compared with its standard error. Without IV estimation, the coefficient is more than twice as large, easily reaching statistical significance.

In section 4, it was demonstrated that using IV estimation produced consistent parameter estimates that were absent when the process was not used. Here we have additionally shown that not using IV estimation can have an impact on the importance placed on each variable.

In table 4, we show the results that would be obtained when a naive approach is taken and neither CMS nor IV estimation takes place. The size of the coefficient for the baseline test is considerably larger than found when CMS and IV estimation takes place (table 2) and even when IV estimation does not take place (table 3). The coefficients for the pupil characteristic and school context variables are also notably different and the substantive conclusions that would be drawn from table 4 are likely to be different from those drawn from the more carefully thought out analysis that brings about table 2.

Table 4: Parameter estimates and standard errors with use of neither CMS nor IV estimation

Coefficient	Model B	Model C	Model D	Model E
Cut-point 1	–2.388(0.080)	–2.381(0.066)	–2.473(0.084)	–2.949(0.140)
Cut-point 2	–0.973(0.070)	–1.179(0.055)	–1.050(0.074)	–1.523(0.133)
Cut-point 3	–0.164(0.069)	–0.506(0.054)	–0.233(0.073)	–0.705(0.132)
Cut-point 4	0.493(0.069)	0.047(0.053)	0.432(0.073)	–0.039(0.132)
Cut-point 5	1.371(0.071)	0.794(0.055)	1.324(0.075)	0.855(0.132)
1st PC for baseline tests	0.453(0.012)		0.447(0.012)	0.450(0.012)
GENDER		0.008(0.032)	–0.155(0.033)	–0.157(0.033)
FSM		0.331(0.036)	0.175(0.037)	0.165(0.037)
CTRDAGE		–0.065(0.005)	–0.021(0.005)	–0.020(0.005)
AMCLANG2		0.056(0.067)	0.290(0.068)	0.286(0.068)
AMCLANG11		–0.673(0.359)	–1.004(0.368)	–0.999(0.368)
AMCLANG12		–0.819(0.466)	–1.052(0.500)	–1.067(0.500)
PCTFSM				0.008(0.003)
AVPCTBASEGT2				0.092(0.020)
School variance	0.477(0.068)	0.202(0.031)	0.460(0.066)	0.371(0.054)

In this paper we have demonstrated the importance and use of both CMS and IV estimation. In particular, we have demonstrated the dangers of ignoring the issues that CMS and IV estimation address – the results obtained look sensible, but can be misleading and there is the possibility of drawing false conclusions.

The MLwiN macro files used to carry out the CMS and IV estimation are available online via http://www.herts.ac.uk/business/staff_public/nhspencer_public/research. Appendix 2 gives details of their implementation.

Appendix 1: Instrumental Variable Estimation

The original independent variables in the model are contained in the matrix ***X***. Variables that act as a set of instrumental variables are contained in a matrix ***Z***. The instrumental variable estimates are obtained using the following equation where $\hat{\beta}$ is the vector of estimated coefficients and *y* is the vector of responses.

$$\hat{\beta} = \left(\boldsymbol{Z}^T \boldsymbol{X}\right)^{-1} \boldsymbol{Z}^T y$$

With Ω being the covariance matrix associated with the residuals, the covariance matrix associated with this estimator is the following.

$$\Sigma_{\beta} = \left(\boldsymbol{Z}^T \boldsymbol{X}\right)^{-1} \boldsymbol{Z}^T \Omega \boldsymbol{Z} \left(\boldsymbol{X}^T \boldsymbol{Z}\right)^{-1}$$

In the practice used in this paper, the instrument set, ***Z***, is identical to the original regressors, ***X***, apart from where the endogenous variable has been replaced with an instrument.

Appendix 2: Use of Macros

The use of the standard MULTICAT macros for MLwiN is described in Yang et al (2001). The adapted version of the MULTICAT macros to allow the use of the probit link function (but not allowing CMS or IV estimation) are available from the authors of this paper. This adapted version, packaged together with the additions for the CMS and IV estimation are available at http://www.herts.ac.uk/business/staff public/nhspencer public/research. In this appendix, we explain how to use the CMS, IV and probit additions to the MULTICAT macros.

To use the IV estimation, we must tell MLwiN what is contained in the instrument set, ***Z*** (see appendix 1 for further details of the IV estimation process). To do this, the macros use a variable “IVSET” which contains a list of the variables in the instrument set, identified by their column number in MLwiN. Thus, if ***Z*** is to contain the variables in columns C3, C5, C6 and C7, then “IVSET” will contain the four digits 3, 5, 6 and 7. With the macro path in MLwiN set to point at the location of the MULTICAT macros with the CMS and IV additions, the prefile should be set to “PRE.IVM” and the postfile to “POST.IVM”. The macros offer the facility to turn the use of IV estimation on or off, so the constant B31 (which controls the switch) should be set to 1 for IV estimation to take place. The current setting of B31 and other option switches can be found by using the command “OBEY OPTIONS.MC”. With this set-up, using MLwiN with its IGLS or RIGLS estimation method will produce an IV estimates of the fixed parameters in the model.

Using the CMS estimation is a two-step process. Firstly, the baseline model with just cutpoints must be fitted and MLwiN is told to store information from this model so that it can be used to scale subsequent models. This is done by using the command “OBEY CMSBASE.MC” once the cut-points only model has been fitted. Secondly, once the subsequent models have been fitted, MLwiN must be told to scale them. This is done by using the command “OBEY CMSSCALE.MC” once the subsequent model has been fitted. The fixed and random coefficients that are then available using the “FIXE” and “RAND” commands are those that have been scaled. It should be noted that in the subsequent models, the variables defining the cut-points should be the first listed in the model.

MLwiN can be set up to use the probit link function by setting the B13 switch to 2 (value 0 corresponds to the logit link and value 1 corresponds to the complementary log-log link).

As above, the current setting of B13 and other option switches can be found by using the command "OBEY OPTIONS.MC".

Although the macros available through the web site detailed above have the CMS and IV estimation macros and a probit link packaged with the standard MULTICAT macros, they can still be used to undertake analyses available with the standard MULTICAT macros. The CMS estimation can be avoided by simply not invoking the "CMSBASE" and "CMSSCALE" macros, the IV estimation can be avoided by setting the B31 switch to zero, and a link function other than probit can be used by setting the B13 switch, as detailed above.

References

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

Fielding, A. (1999) Why use arbitrary points scores: ordered categories in models of educational progress. *Journal of the Royal Statistical Society, Series A*, 162, 3, pp 303-328.

Fielding, A. (2003) Scaling for residual variance components of ordered category responses in generalised linear mixed multilevel models. *Quality and Quantity*, 38, 4, pp 425-433.

Fielding, A. (2002) Ordered category responses and random effects in multilevel and other complex structures: scored and generalised linear models *in Multilevel Modeling: Methodological Advances, Issues and Applications* (S. Reise & N. Duan, eds.), Erlbaum: New Jersey.

Fielding, A. & Yang, M. (2005) Generalised linear mixed models for ordered responses in complex multilevel structures: effects beneath the school or college in education. *Journal of the Royal Statistical Society, Series A* (to appear).

Snijders, T.A.B. & Bosker, R.J. (1999) *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, Sage Publications: London.

Spencer, N. H. (2002) Combining modelling strategies to analyse teaching styles data. *Quality and Quantity*, 36, 2, pp 113-127.

Spencer, N.H. (2003) Consistent parameter estimation for lagged multilevel models. *Journal of Data Science*, 1, 2, pp 123-147.

Spencer, N. H. & Fielding, A. (2000) An instrumental variable consistent estimation procedure to overcome the problem of endogenous variables in multilevel models. *Multilevel Modelling Newsletter*, 12, 1, pp 4-7.

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

Yang, M., Rasbash, J., Goldstein, H. & Barbosa, M. (2001) *MLwiN Macros for Advanced Multilevel Modelling, V2.0a*. Multilevel Models Project, Institute of Education, University of London.