H. Ernste (ed.). (1996) Multilevel analysis with structural equation models. Zürich: ETH.

2. A comparison of some multilevel sibling models

J.J. Hox[1], *J. Dronkers*[1], *M. Evans*[2], *J. Kelley*[2]

2.1 Introduction

A favored tool for the analysis of the effects of parental background on the educational attainment of their children is structural equation analysis using a sibling model. The data contains information on the educational attainment of two or more brothers or sisters from the same family and background information on their parents. The effect of shared environmental and genetic factors is visible in the covariation of the siblings' educational attainment. Sibling models attempt to separate the common family effects on this covariation from the unique effects on the individual education. Parental education and occupational level are often used as explanatory variables, but unmeasured common environmental or genetic characteristics and reciprocal influences are also possible explanations for the effect of the common family background on the educational attainment.

A structural equations model to distinguish between common family and individual effects was developed by Hauser and Wong (1989). This model is a MIMIC model in which the latent variable represents the common educational family factor. This model can be extended to include individual explanatory variables.

A different approach to the analysis of family effects on educational attainment is multilevel analysis, which considers the children as nested within families. Two different multilevel models are discussed here: a multilevel regression model and a multilevel structural model. Multilevel regression models (also known as hierarchical regression models and random coefficient models) predict a dependent variable, measured at the lowest (individual) level, using explanatory variables at all available levels. Multilevel structural

[1] Faculty of Educational Science, University of Amsterdam, Wibautstraat 4, NL-1091 GM Amsterdam, the Netherlands, tel: +31-20-525 12 01, fax: +31-20-525 15 00, e-mail: hox@educ.uva.nl resp. jaapd@educ.uva.nl

[2] Research School of Soocial Sciences. Australian National University, Canberra ACT, Australia, fax: +61-62-97 29 37, e-mail: mariah@coombs.anu.edu.au

models are structural equation models that jointly analyze both individual level and group level covariation.

This chapter compares the Hauser/Wong model with multilevel regression and multilevel structural models. The goal is to examine how much the models really differ: can different models be translated into each other, and if they can, do they produce similar estimates for corresponding parameters? To answer these questions, models are formulated for sibling data from the Australian National Social Science Survey. Since our goal is comparing models, prototypic models are used with only a small number of variables. A more extensive analysis of these data using the Hauser/Wong model is reported in Borgers et al. (1995).

2.2 Data

The data are from the Australian Social Science Survey 1989-1990. They contain information on the educational attainment and background variables of the respondents and up to three siblings. Included in the analysis are all respondents elder than 18 years with at least one sibling who was also elder than 18 years. This results in data from 10795 individuals in 3706 families (an average of 2.9 siblings per family, the range is from 1 to 14 siblings). Since information was asked for up to three siblings, we may have information on 1-4 individuals per family.

The family variables used in the analyses are the father's educational attainment (*FEDU*), father's occupational status (*FOCC*), mother's educational attainment (*MEDU*), and the number of siblings (*NSIB*). The individual variables are the position (oldest/youngest: *O/Y*) and the sibling's educational attainment (*SED*).

2.3 Models

2.3.1 The Hauser/Wong model

In this analysis, the unit of analysis is not the individual, but pairs of siblings. For each family, all possible pairs of siblings are formed, and in each pair a distinction is made between the oldest and the youngest of the two siblings. Since large families contribute more sibling pairs than small families, each pair is weighted by the inverse of the number of pairs from its family. Thus, all families are represented equally in the analysis.

The basic model by Hauser and Wong (1989) is a MIMIC model, displayed in Figure 2.1. The two siblings are represented by two distinct variables: *OSED* (Oldest Sibling EDucation) and *YSED* (Youngest Sibling EDucation). The common educational factor is indicated by the educational level of

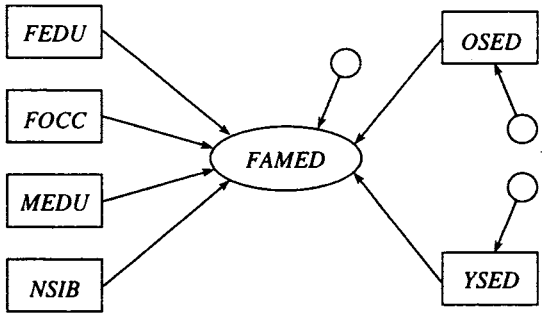

Figure 2.1: Basic model by Hauser & Wong (1989) for class and family effects

both siblings. This common education is influenced by the family background variables. The model can be extended by including more family or individual variables. Additional family variables would all be modeled by having an effect on the common education factor. Additional individual variables would be modeled by specifying an effect directly on the education of the siblings (e.g., the sibling's gender would be modeled by including one gender variable for the oldest and one for the youngest sibling, and specifying a path from each gender variable to the corresponding sibling education variable). Such extensions are straightforward, and will not be discussed here.

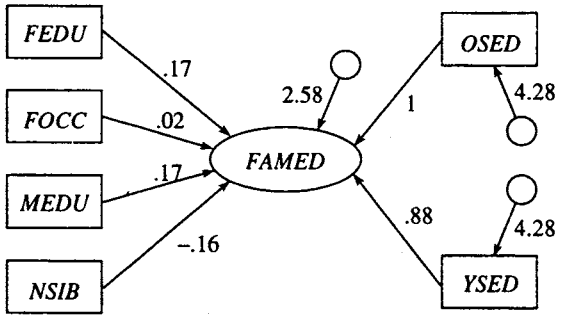

Figure 2.2: Estimated parameters for basic Hauser/Wong model (Borgers et al., 1995)

Figure 2.2 presents the basic model with the parameter estimates (ML solution, estimates taken from Borgers et al., 1995). This model explains

41% of the variation of the common education factor, and 50%/44% of the educational level of the oldest/youngest sibling.

For identification, the path of the common education factor to *OSED*, the oldest sibling's education, is fixed at 1. In stratification research, the factor loading of *YSED*, the youngest sibling's education, is interpreted as the proportional adjustment of the family's characteristics on the youngest sibling. As such, it is an indicator for the degree of similarity of educational level within families. In these data, this factor loading is estimated as 0.88 with a standard error of 0.03; this makes it clearly different from 1. The effect of the family background variables is attenuated for younger siblings.

From a statistical viewpoint, the difference between the estimated loading of *YSED* and the loading of *OSED*, which is fixed at 1, models an *interaction* between the family characteristics and the position of the siblings. For instance, the total effect of the father's education on the sibling's education can be calculated by multiplying all unstandardized path coefficients on the path (cf. Bollen, 1989). For example, the effect of *FEDU* on *OSED* is $0.17 * 1 = 0.17$, and the effect of *FEDU* on *YSED* is $0.17 * 0.88 = 0.15$.

There are two peculiarities in the way the Hauser/Wong model handles the interaction between the family variables and the sibling's position. First, because all family effects are channeled through the common latent variable, the proportional adjustment will be the same for all family variables. Since the interaction effects might well differ for different family variables, this amounts to an assumption that is not tested in the model. Second, if there are more than two siblings, it is not clear what the adjustment really means. As noted above, for families with more than two siblings all possible sibling pairs are formed and entered in the analysis. Thus, if there are three siblings, in order of age called Oldest, Middle, and Youngest, the adjustment for all pairs is estimated as 0.88. However, if the adjustment for the pair Oldest-Middle is 0.88, and for Middle-Youngest is also 0.88, the total adjustment for Oldest-Youngest should be $0.88 * 0.88 = 0.77$, which is of course not identical to 0.88. Thus, if there are more than two siblings, the value of the adjustment has no simple interpretation.

2.3.2 The multilevel regression model

In this analysis, the units of analysis are both the families and the individual siblings. The model is a regression model, with explanatory variables at both the family and the individual level. There are residual errors for both levels. Also, the regression coefficients for the individual level predictors may vary between families, and this variation can be modeled by the family level variables. Technically, this is accomplished by introducing cross-level interactions: interactions between the individual level predictors and the family level predictors. For a technical discussion of the multilevel regression model see Bryk and Raudenbush, 1992), a nontechnical introduction is given by Hox (1995).

Table 2.1: Results multilevel regressions

variables	model				
	null	fixed ind.lvl	random slope	family level	cross-lvl interaction
const	10.86	10.86	10.86	7.96	9.47
eldest		$-.00^{ns}$	$-.00^{ns}$	$-.04^{ns}$	$-.04^{ns}$
fedu				.13	.13
focc				.02	.02
medu				.17	.17
nsib				−.15	−.15
*eldest*medu*					.07
σ^2	4.01	4.01	3.79	3.80	3.80
σ^2_{const}	4.11	4.11	3.80	2.37	2.37
σ^2_{eldest}			.68	.64	.60
deviance	50588	50588	50538	49297	49273

In our case, we have the individual level variable Eldest (a dummy variable indicating whether an individual is the eldest sibling or not), and the family level variables *FEDU* (Father EDUcation), *FOCC* (Father OCCupation), *MEDU* (Mother EDUcation), and *NSIB* (Number of SIBlings). Preliminary analyses showed that the variance of the dependent variable *SED* (Siblings EDucation) is decomposed in an individual level variance of 4.11 and a family level variance of 4.01. In other words, 51% of the variance is at the individual level, and 49% of the variance is at the family level. Also, the regression coefficient of the individual level variable Eldest does vary between families, a variation that can (partly) be explained by including a cross-level interaction with the family level variable *MEDU*. The final model explains 5% of the individual level variance and 42% of the family level variance. In total, 24% of the variance is explained. Table 2.1 presents the results of the multilevel regression analysis (full Maximum Likelihood solution using MLn, Rasbash & Woodhouse, 1995) for a succession of models (cf. Hox, 1995). A path representation of the final model is presented in Figure 2.3. To simplify interpretation, the interaction variable is calculated using centered values for *MEDUC* and Eldest.

2.3.3 The multilevel structural equations model

A convenient approach towards structural equations modeling for multilevel data is outlined by Muthén (1994) (see also McDonald, 1994, and Hox, 1995). In this approach, the individual level variables are decomposed into a family mean and the individual deviation from that mean. This leads to a between families (family level) covariance matrix $\mathbf{S}_B$ and a within families (individual level) covariance matrix $\mathbf{S}_W$. Multilevel structural models are formulated and

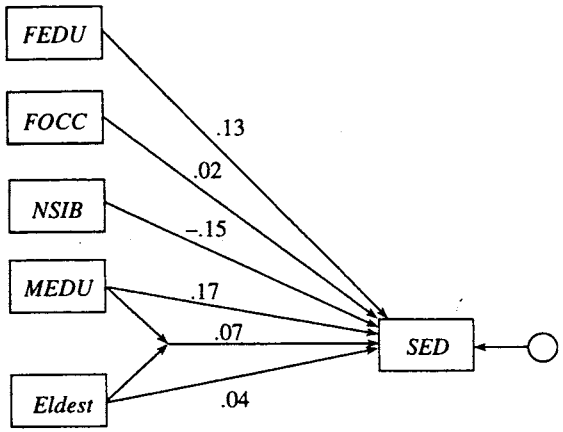

Figure 2.3: Estimates multilevel regression model with cross-level interactions

estimated using the multigroup option of conventional software. The model for Σ_W is specified for both $\mathbf{S}_W$ and $\mathbf{S}_B$, with equality restrictions between these two 'groups', and the model for Σ_B is specified for $\mathbf{S}_B$ with a scale correction (for details see Muthén, 1994, or Hox, 1995). If all families have the same size, the approach leads to a full Maximum Likelihood solution. If the families have different sizes, a limited information solution is obtained. Simulations by McDonald (1994) and Hox (1993) suggest that the limited information solution is sufficiently accurate for practical applications.

A multilevel path model for the example data is presented in Figure 2.4. In this figure, the model on the left is the within families (individual level) model; it is simply a regression model predicting the individual education with the individual level variable Eldest. The model on the right is the between families (family level) model. It specifies the within family model with equality constraints between the two 'groups', and a family level regression model for *SED* (Sibling EDucation). No attempt is made to model the variable 'Eldest' at the family level; it is treated as an exogenous model for which only a family level variance is specified.

The multilevel path model in Figure 2.5 did not fit very well ($\chi^2_5 = 506$). Adding just one covariance term between the family level variable *NSIB* and the family level variable Eldest leads to a model with a satisfactory fit ($\chi^2_4 = 9.5$, $p = 0.05$). This extra covariance term is acceptable. At the family level the variable Eldest represents the proportion of eldest children in the family, which is obviously a function of the family size. Thus, the extra covariance can be viewed as a nuisance term, and left uninterpreted. Figure 2.5 presents the between family path diagram (which includes the

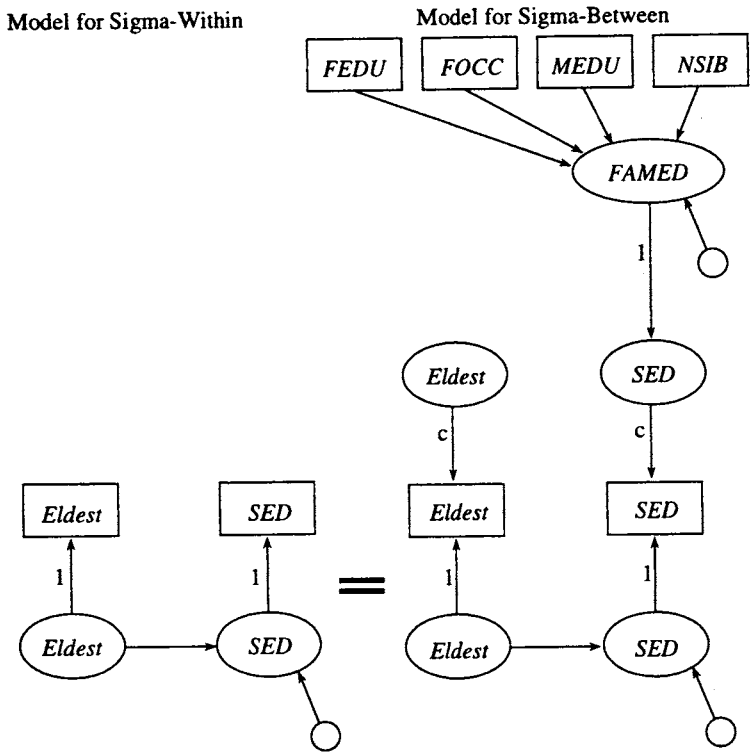

Figure 2.4: Multilevel structural model using multiple group specification

within family model as well) with the limited information estimates (obtained by maximum likelihood estimation of the unbalanced matrices). To simplify interpretation, the covariances among the exogenous variables *FEDU*, *FOCC*, *MEDU*, and *NSIB* are omitted, as is the added covariance term between *NSIB* and *GELDEST*.

2.4 Model comparisons and discussion

It is instructive to compare the various parameter estimates of the three models. This requires some care, since the variables do not have the same meaning on all levels. As discussed above, the Hauser/Wong model effectively models an equal interaction effect for all family level explanatory variables.

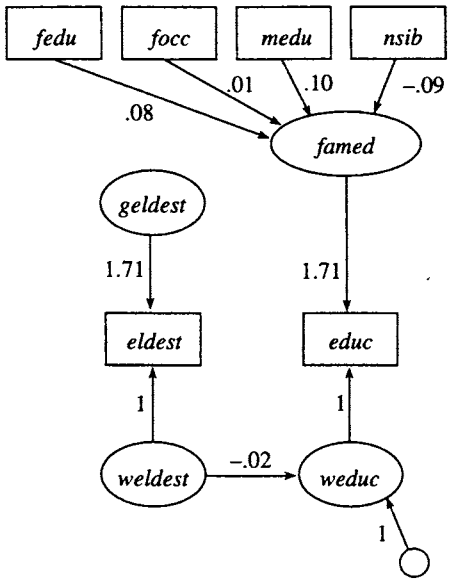

Figure 2.5: Estimates final multilevel structural model

The total effect for these variables can be found by multiplying the (unstandardized) coefficients on the path. The multilevel regression model specifies just one interaction, for *MEDU* and Eldest. Both models can be compared by tabulating the regression coefficients for the eldest and the youngest siblings separately (using unstandardized and uncentered coefficients). The multilevel structural model is different, because it involves two separate models for the individual and the family level. Also, a number of (fixed) scaling coefficients are involved that must be taken into account. Table 2.2 shows the results for the family variables.

The corresponding parameter estimates are similar. The interaction in the data appears to be modeled most effectively by the multilevel regression model, since it allows precise specification of interactions. It also shows (cf. Table 2.1) that the specified interaction explains only a small part (11%) of the variation in the slopes of Eldest between families. In the Hauser/Wong model this interaction appears to be much smaller because it is shared among the four family level variables *FEDU*, *FOCC*, *MEDU* and *NSIB*, while only one of these four vaiables actually shows a significant interaction effect. The multilevel structural model does not specify interactions. Specifying interactions in structural equation models is a complicated procedure, which is not discussed here.

Table 2.2: Total effects on of family variables on sibling education in the three models

variables	model				
	Hauser/Wong		multilevel regression		multilevel SEM
	eldest	youngest	eldest	youngest	family level
FEDU	.17	.15	.13	.13	.14
FOCC	.016	.014	.017	.017	.017
MEDU	.17	.15	.22	.15	.17
NSIB	−.16	−.14	−.15	−.15	−.15

The three models deal quite differently with the effect of the individual variable Eldest. The standard Hauser/Wong model investigates the difference between the effects of the family level variables (conceptualized as a single latent factor) on the eldest and the youngest sibling. It does not estimate the mean difference between the eldest and youngest sibling. However, this can be accomplished easily by adding means to the factor model. As discussed earlier, with data on more than two siblings, the Hauser/Wong model leads to a peculiar and inconsistent definition of the variable Eldest. This problem does not occur in the multilevel regression and the multilevel structural model. In these models, Eldest is unequivocally coded to indicate whether an individual is the eldest or not (since we have information on a siblings' position in its family, a more detailed operationalization could be attempted in these models). Both models include the effect of Eldest on the siblings' education. In both models the effect is small (.04 in the regression model, .02 in the structural model) and insignificant. However, the multilevel structural model conceals the between family variation of the slopes of Eldest.

When path coefficients are made comparable (which requires some hand-calculations) the three models show similar results. The Hauser/Wong model and the multilevel regression model have a comparable structure. The multilevel regression model holds an advantage, because it can model more specific interactions and test the residual variation of the slopes of the individual variables. The Hauser/Wong model can be used to test for specific constraints, such as equality of the factor loadings of the two siblings, but similar constraints are possible in the multilevel regression framework (cf. Bryk & Raudenbush, 1992). If we wish to construct more complex path models, including individual or family level intervening variables, we need a full structural equations model. Hauser has developed models that include both exogenous and endogenous family variables (cf. Cuttance & Ecob, 1988). Especially when there are only two siblings, this is an attractive approach. If there are more than two siblings, a multilevel approach is more appropriate. However, current procedures do not include random coefficient path models. If multilevel structural models are used, it appears prudent to use a series of multilevel regression models first to search for varying coefficients and cross-

level interactions. If these are not negligible, it may be possible to include them in the multilevel structural model.

2.5 References

Bollen, K. (1989) *Structural Equations with Latent Variables.* Wiley, New York.

Borgers, N., Dronkers, J., Rollenberg, L., Evans, M. & Kelley, J. (1995) *Educational Resemblance between Australian Siblings: Differences of Gender, Generations, Migration, Family Forms and Mothers' Work.* Paper presented at the annual conference of the American Sociological Association, August 19-23, Washington DC.

Bryk, A.S. & Raudenbush, S.W. (1992) *Hierarchical Linear Models.* Sage, Newbury Park.

Cuttance, P.F. & Ecob, R. (1988) *Structural Modeling by Example: Applications in Educational, Sociological and Behavioral Research.* Cambridge University Press, New York.

Hauser, R.M. & Wong, R.S.-K. (1989) Sibling resemblance and intersibling effects in educational attainment. *Sociology of Education*, Vol. 62, pp. 149-171.

Hox, J.J. (1993) Factor analysis of multilevel data. In: Oud, J.H.L. & van Blokland-Vogelesang, R.A.W. (eds.) *Advances in Longitudinal and Multivariate Analysis in the Behavioural Sciences.* ITS, Nijmegen.

Hox, J.J. (1995) *Applied Multilevel Analysis.* Thesis, Amsterdam.

McDonald, R.P. (1994) The bilevel reticular action model for path analysis with latent variables. *Sociological Methods & Research.* Vol. 22, pp. 399-413.

Muthén, B. (1994) Multilevel covariance structure analysis. *Sociological Methods & Research.* Vol. 22, pp. 376-398.

Rasbash, J. & Woodhouse, G. (1995) *MLn Command Reference.* Multilevel Models Project, University of London, London.