Demographic Research a free, expedited, online journal
of peer-reviewed research and commentary
in the population sciences published by the
Max Planck Institute for Demographic Research
Konrad-Zuse Str. 1, D-18057 Rostock · GERMANY
www.demographic-research.org

DEMOGRAPHIC RESEARCH
VOLUME 9, ARTICLE 2, PAGES 25-40
PUBLISHED 10 SEPTEMBER 2003
www.demographic-research.org/Volumes/Vol9/2/
DOI: 10.4054/DemRes.2003.9.2

Reflexion

The Problematic Estimation of "Imitation Effects" in Multilevel Models

Øystein Kravdal

Table of Contents

1	Introduction	26
2	Theoretical Issues	27
3	Simulation Experiments	29
3.1	Data and Variables	29
3.2	A Continuous Model that Includes the Average Outcome	31
3.3	Other Continuous Models	33
3.4	How Can the “Imitation Effect” be Estimated?	33
3.5	Logistic Models	34
4	Conclusion	36
5	Acknowledgements	36
	Notes	37
	References	40

Reflexion

The Problematic Estimation of "Imitation Effects" in Multilevel Models

Øystein Kravdal [1]

Abstract

It seems plausible that a person's demographic behaviour may be influenced by that among other people in the community, for example because of an inclination to imitate. When estimating multilevel models from clustered individual data, some investigators might perhaps feel tempted to try to capture this effect by simply including on the right-hand side the average of the dependent variable, constructed by aggregation within the clusters. However, such modelling must be avoided. According to simulation experiments based on real fertility data from India, the estimated effect of this obviously endogenous variable can be very different from the true effect. Also the other community effect estimates can be strongly biased. An "imitation effect" can only be estimated under very special assumptions that in practice will be hard to defend.

[1] Department of Economics, University of Oslo. E-mail: oystein.kravdal@econ.uio.no

1. Introduction

Many demographic studies are now based on a multilevel approach, for which special statistical techniques have been developed over the last two decades (e.g. Goldstein 1995). Using such an approach means that the investigator takes into account, for example, that an individual's behaviour not only depends on the resources and attitudes of that individual, but also community factors (see e.g. Blalock (1984) for a general, non-technical discussion of some potentials and problems related to multilevel modelling). The community factors may be characteristics of other individuals in the community, which can be a geographically defined area or a socially defined reference group, or institutional or other so-called "global" factors with no individual-level counterpart. In addition, it is plausible that individual behaviour is partly determined by whether other people in the community behave similarly. In fact, it has been argued that this may be the main causal channel between other people's characteristics and the individual behaviour (Erbring and Young 1979). To estimate such an "imitation" or "conformity effect", one might be tempted to include among the community factors the average of the outcome variable in focus or perhaps a measure of its distribution. In that case, one approach would be to use a survey where the respondents' community of residence is identified and find data for these communities in another data source, for example a census. Unfortunately, this is often not possible. As an alternative, one may use data that include clusters of individuals reported to be living in the same area, and construct various aggregate variables for these areas, to be linked with the original individual file. This could be done, for example, with the Demographic and Health Surveys, which are frequently used to analyse fertility and child mortality in developing countries. There may be few respondents in each area, but that itself is not necessarily a problem (see below).

The latter approach was taken by McNay, Arokiasamy and Cassen (2003), who estimated multilevel logistic models for the probability of using contraception among uneducated women in India. They included a categorical variable for the proportion using contraception among all interviewed women in the district, regardless of these women's education. When this "imitation effect" was taken into account, the effect of the literacy rate in the district was no longer significant, and the district-level variation was strongly reduced.

While being theoretically appealing, the community average of the outcome among the respondents in the sample is an endogenous variable that is correlated with the unobserved factors of importance for the individual outcome. It is common knowledge that one should be careful to include such variables that are correlated with the error term, although all demographers are perhaps not fully aware of how strongly biased the estimates may actually be. The objective of this paper is to illustrate by some

simulation experiments that one may arrive at wildly wrong conclusions if the temptation to include this type of variables is not resisted. Whereas this may be a trivial warning to those who are most strongly inclined to shy away from endogenous variables, it may be helpful to others. After all, we often include exogenous variables that we expect to be linked with various unobserved factors, accompanied by a discussion of the limitations of the estimates, and it may seem a small step to also include, say, an average of the dependent variable. Leaning heavily on Manksi (1993), it is also explained in the paper how an "imitation effect", in principle, might be estimated, after rearranging the equation.

I first consider continuous models, which are easiest to handle mathematically. After reviewing some mathematical arguments, I show the results from a simulation experiment with a continuous model that includes the average outcome. Subsequently, I consider two other measures of how common the outcome is in the community. Finally, I turn to logistic models, including one that is similar to that estimated by McNay et al. For simplicity, the focus is on models with an error term that is not split into an individual- and a community-level contribution.

2. Theoretical Issues

Let us first consider the model

$$(1)\quad y_{ik} = a_0 + a_1 y_{\cdot k} + a_2 x_{ik} + a_3 x_{\cdot k} + e_{ik}$$

where y_{ik} is a characteristic of person i in community k, x_{ik} another individual characteristic, $y_{\cdot k}$ and $x_{\cdot k}$ the corresponding averages over all individuals in the community, and e_{ik} an individual-level normally distributed error term with mean 0 that is uncorrelated with x. a_0, a_1, a_2 and a_3 are effect parameters. This model is a special case of that discussed by Manski (1993).

The variable $y_{\cdot k}$ is obviously correlated with the error term. Taking the average

$$(2)\quad y_{\cdot k} = a_0 + a_1 y_{\cdot k} + (a_2 + a_3)\, x_{\cdot k} + e_{\cdot k}$$

and rearranging gives

$$(3)\quad y_{\cdot k} = a_0/(1-a_1) + (a_2 + a_3)\, x_{\cdot k}/(1-a_1) + e_{\cdot k}/(1-a_1)$$

from which it is easy to see that the covariance between $y_{\cdot k}$ and e_{ik} is proportional with the variance of e_{ik} . Because of this correlation, one will get a biased estimate of a_1 if,

for example, $y_{\cdot k}$, x_{ik} and $x_{\cdot k}$ are fed into an OLS estimation module as if they were ordinary regressors uncorrelated with the error term. This is further dealt with below.

To circumvent the problem arising from the correlation between a regressor and the error term, (3) can be inserted into (1). This gives

$$(4) \quad y_{ik} = a_0 /(1-a_1) + (a_3 + a_1a_2)\, x_{\cdot k}/(1-a_1) + a_2x_{ik} + a_1\, e_{\cdot k}/(1-a_1) + e_{ik}$$

If the regressors are linearly independent, the coefficients $a_0 /(1-a_1)$, $(a_3 + a_1a_2) /(1-a_1)$ and a_2 can be identified. However, this is not sufficient to identify all coefficients a_0, a_1, a_2 and a_3.

As pointed out by Manski (1993), it helps to assume that there is no effect of the community variable $x_{\cdot k}$, i.e. that the model is

$$(5) \quad y_{ik} = a_0 + a_1y_{\cdot k} + a_2x_{ik} + e_{ik}$$

Solving for $y_{\cdot k}$ and inserting, as above, yields

$$(6) \quad y_{ik} = a_0 /(1-a_1) + a_1a_2\, x_{\cdot k}/(1-a_1) + a_2x_{ik} + a_1\, e_{\cdot k}/(1-a_1) + e_{ik}$$

which is sufficient to estimate all coefficients.

In fact, the model is in principle identified (assuming no linear dependence between the involved regressors) if there is one individual variable for which there is no corresponding community variable. (This particular point is not made by Manski). Another example of such a model is:

$$(7) \quad y_{ik} = a_0 + a_1y_{\cdot k} + a_2x_{ik} + a_3\, x_{\cdot k} + a_4z_{ik} + a_5\, q_{\cdot k} + e_{ik}$$

where z_{ik} is another individual-level variable and $q_{\cdot k}$ is a community-level variable. In this case, the equation corresponding to (4) is

$$(8) \quad y_{ik} = a_0 /(1-a_1) + (a_3 + a_1a_2)\, x_{\cdot k}/(1-a_1) + a_2x_{ik} + a_1a_4\, z_{\cdot k}/(1-a_1) + a_4z_{ik} + a_5\, q_{\cdot k} /(1-a_1) + a_1\, e_{\cdot k}/(1-a_1) + e_{ik}$$

where a_4 is readily estimated and the ratio $a_1a_4 /(1-a_1)$ can be used to find a point estimate of a_1. An approximate measure of the standard error can be used for statistical inference.

If the model includes *two or more* individual-level variables that have no community-level counterparts, we run into an "over-identification" problem, because there is more than one ratio that can be used to estimate a_1.

Let us now return to the simpler model (5). If an OLS estimation is tried with $y_{\cdot k}$ and x_{ik} as regressors, the a_1 effect will be biased because of the correlation between $y_{\cdot k}$ and the error term, as explained above. It can be shown (Note 1) that the bias is

(9) $\quad \text{Corr}(y_{\cdot k}, e_{ik})\ \text{Std}\ e_{ik} / (\text{Std}\ y_{\cdot k}\ (1-\text{Corr}^2(y_{\cdot k}, x_{ik}))$.

As the error term increases, $y_{\cdot k}$ is more and more "dominated" by $e_{\cdot k} / (1-a_1)$ and less and less correlated with x. Thus, the expression in (9) approaches $1 - a_1$. In other words, one will not estimate the true a_1 effect, but $a_1 + 1 - a_1$, which is 1.

With a model such as (1), OLS will give a a_1 estimate of 1 regardless of the size of the error term. That is because of the identification problem noted above, which only "disappears" with a_1 assumed to be 1, in which case $\hat{a}_0 = 0$ and $\hat{a}_2 = -\hat{a}_3$, with $\hat{a}_2$ being determined from the intra-cluster variation. (To see this, set a_1 to 1 and subtract $y_{\cdot k}$ from both sides of (2), which gives $0 = (a_2+a_3)x_{\cdot k} + e_{\cdot k}$) Put differently, there is one trivial solution with $a_1=1$ and infinitely many others.

3. Simulation Experiments

I now turn to simulations to illustrate the size of the bias introduced when estimating (1), (5), (7) or similar models directly, without first getting rid of the average-outcome variable.

3.1. Data and Variables

The simulations are based on real data from the Indian National Family Health Survey of 1998-99 (International Institute for Population Sciences and ORC Macro, 2000) and realistic parameters derived from estimation of similar models based on these data. The survey has a clustered sample of about 90000 women who live in more than 3000 census-enumeration areas, each of which spans one or a few villages or part of a town or city.

I do not include all women in the survey, but only the 24278 women who were 30-50 years old at the time of interview, had less than five years of education, and lived in census-enumeration areas where there was also at least one woman who had five or more years of education. The latter restriction allows two types of averages to be constructed, one for the women under analysis and one for the better educated in the same area. This sub-sample includes 2495 census-enumeration areas and about 10 women in each area. These women are referred to below as a cluster. Weights that can

be used to make the sample nationally representative are available in the data, but ignored in this study, for simplicity.

An outcome is simulated for all these 24278 women, using a model that includes the cluster average of the outcome variable or a similar measure (not directly observable; see below) and a few other individual and community characteristics. Realistic effect parameters are found by first estimating a model that includes all these characteristics along with the average-outcome variable for the interviewed women in the same census-enumeration area who have five or more years of education, and who are not part of the sample under analysis. This reflects an idea that the effect of the behaviour among other women with less than five years of education perhaps is not very different from the effect of the behaviour among the better educated in the same community (and that the latter is adequately estimated, i.e. not picking up other community factors, as discussed below). Fortunately, this is not a critical assumption. The crucial issue from the perspective of this study is whether the effect parameter that is used in the simulation is the same as the effect that is estimated from these simulated outcomes, and the conclusion about that is not sensitive to the choice of parameter.

In the continuous models that I consider, the outcome variable is the number of live births up to interview, which ranges from 0 to 15, with mean value 4.5 and standard deviation 2.2. The following variables are included in these models: The woman's own education (in years), her age (in years), the average number of years of education in the census-enumeration area, and a dummy that is set to 1 if the area is reckoned as urban (0 if rural). The average educational level was found by Kravdal (2002) and Moursund and Kravdal (2003) to have a large effect on fertility, net of the woman's own education. These studies also showed that it was unproblematic to use a community-education variable based on so few people as in the DHS clusters.

The simulations yield, of course, non-integer and some negative values of the number of children, but this should be of no concern, given the objective of the study. It should also be noted that causal interpretation of estimates is particularly difficult because the outcome variable reflects events up to 35 years earlier, whereas the included variables refer to the situation at interview. However, this is not an important limitation of this purely methodological contribution.

Logistic models for the probability of having at least three children are also estimated. The same variables are included in these models.

The simulation and estimation are done in SAS.

3.2. A Continuous Model that Includes the Average Outcome

Let us first assume that the number of children y_{ik} born to a woman i in community k is given by:

$$(7) \quad y_{ik} = a_0 + a_1 y_{\cdot k} + a_2 x_{ik} + a_3 x_{\cdot k} + a_4 z_{ik} + a_5 q_{\cdot k} + e_{ik}$$

where e_{ik} is an individual-level normally distributed error term with mean 0 that is uncorrelated with x, z and q. x_{ik} is the woman's education, $x_{\cdot k}$ the corresponding cluster average, and z_{ik} is her age. $q_{\cdot k}$ is a rural/urban dummy, defined at the community level without any individual-level counterpart.

An OLS model with regressors $y^{+}_{\cdot k}$, x_{ik} , $x_{\cdot k}$, z_{ik} and $q_{\cdot k}$ is first estimated. $y^{+}_{\cdot k}$ is the average number of children among women with five or more years of education in the census-enumeration area, and is included instead of $y_{\cdot k}$. The estimates are

$\hat{a}_0^{\,0}$=0.41, $\hat{a}^{+}_{1}{}^{0}$ = 0.26, $\hat{a}_2^{\,0}$= -0.05, $\hat{a}_3^{\,0}$=-0.36, $\hat{a}_4^{\,0}$ = 0.09, and $\hat{a}_5^{\,0}$ = -0.07.

All are significantly different from 0 at the 0.05 level. The standard deviation of e is estimated to be 2.06.

The parameters in the simulation model are set to these estimates:

a_0 = 0.410, a_1 = 0.260, a_2 = -0.050, a_3 = -0.360, a_4 = 0.090 , and a_5 = -0.070,

and e_{ik} is drawn for each individual, using a normal distribution with standard deviation 2.06. The average $y_{\cdot k}$ of the simulated outcomes is, of course, not directly observable, but can be calculated by taking the average and rearranging (in analogy with (2), (3)) and afterwards insert into (7). (This is, of course, the same as feeding the parameters directly into (8).) After this procedure, a sample that satisfies (7) is established. The error term is uncorrelated with x, z and q (the calculated correlation coefficients are about 0.004), but, of course, not with $y_{\cdot k}$ (correlation coefficient is 0.30). The predicted number of children ranges from –3.3 to 13.7. The average value is 4.9 and the standard deviation is 2.2, which are close to those observed in the data.

OLS regression based on the simulated outcomes gives:

$\hat{a}_0$=-2.820, $\hat{a}_1$ = 0.957, $\hat{a}_2$= -0.046, $\hat{a}_3$=-0.012$^{\#}$, $\hat{a}_4$ = 0.080, and $\hat{a}_5$= -0.013$^{\#}$,

where $^{\#}$ means that the effect is not significant at the 0.05 level. The standard deviation of e_{ik} is estimated to be 1.95.

The effects of the individual variables are thus quite correctly estimated, but not the others. The $\hat{a}_1$ estimate is almost 1, as expected when the correlation with the error term is fairly large, and far from the parameter $a_1 = 0.26$ used in the simulation. When the standard deviation of the error term is instead set to 0.2 or 0.02 (which gives simulated outcomes with much smaller variation than actually observed), $\hat{a}_1$ becomes 0.35 or 0.26, respectively. In the latter case, all estimates are, of course, very close to the true ones. Conversely, the $\hat{a}_1$ estimate is 0.979 when the standard deviation is increased to 3 and 0.998 when it is increased to 10. (Note 2)

It should be noted that the small values of $\hat{a}_3$ and $\hat{a}_5$ are not a necessary consequence of the inclusion of $y_{.k}$. The values can become large with other choices of $q_{.k}$ and $x_{.k}$.

If the model does not include at least one individual-level variable without a community-level counterpart, one runs into an identification problem, as explained above. One example of such a model is:

$$(10) \quad y_{ik} = a_0 + a_1 y_{.k} + a_2 x_{ik} + a_3 x_{.k} + a_4 z_{ik} + a_5 z_{.k} + a_6 q_{.k} + e_{ik}$$

A simpler example is

$$(1) \quad y_{ik} = a_0 + a_1 y_{.k} + a_2 x_{ik} + a_3 x_{.k} + e_{ik}$$

A simulation experiment similar to those described above confirms the theoretical arguments. The effect parameters in (1) are first set to

$a_0 = 3.67$, $a_1 = 0.260$, $a_2 = -0.070$, $a_3 = -0.310$, Std e = 2.11 .

OLS regression based on the simulated outcomes gives:

$\hat{a}_0 = 0.000$, $\hat{a}_1 = 1.000$, $\hat{a}_2 = -0.064$, and $\hat{a}_3 = 0.064$.

The standard deviation of e is estimated to be 1.99. The estimate of a_2 approaches the correct value of –0.070 as the standard deviation of the error term in the simulation is reduced, but $\hat{a}_1$ remains 1, and $\hat{a}_0$ and $\hat{a}_3 + \hat{a}_2$ remain 0.

Similarly, simulation based on model (10) and subsequent estimation gives:

$\hat{a}_0 = 0$, $\hat{a}_1 = 1$, $\hat{a}_2 = -\hat{a}_3$, $\hat{a}_4 = -\hat{a}_5$, $\hat{a}_6 = 0$

3.3. Other Continuous Models

As an alternative, one might consider including the average $y^*_{\cdot k}$ for the entire survey sample in the community, not only those with less than five years of education. The model would then be:

$$(11) \quad y_{ik} = a_0 + a_1 y^*_{\cdot k} + a_2 x_{ik} + a_3 x_{\cdot k} + a_4 z_{ik} + a_5 q_{\cdot k} + e_{ik}, \quad y^*_{\cdot k} = p y_{\cdot k} + (1-p)\, y^+_{\cdot k}$$

where p is the proportion of women with 0-4 years of education in the community, and $y^+_{\cdot k}$ is the average of y_{ik} among the better-educated women. When I assume that the latter average is uncorrelated with the error term and proceed with the simulation and estimation as above (which involves a somewhat more complex equation than (3) to find $y^*_{\cdot k}$), the $\hat{a}_1$ estimate is 0.849. This is still far from the 0.260 used in the simulation, but the difference is not quite as large as in the simulation exercise above, reflecting the weaker correlation between this endogenous variable and the error term. (As the error term increases, the $\hat{a}_1$ estimate approaches 1.34.)

One might also consider excluding the woman in focus before calculating a cluster average of the dependent variable, in analogy with what has been done for exogenous variables (such as $x_{\cdot k}$) in some studies. This would mean that $y_{\cdot k}$ is substituted with $y_{\cdot k}^{(i)}$, defined as $(y_{\cdot k}\, n_k - y_{ik})/(n_k-1)$, where n_k is the number of women in the cluster. (Note 3). Although the person's own outcome is disregarded when forming such an average, and the variable in that sense is "less obviously" endogenous, there is still a correlation with the error term. For example, the average $y_{\cdot k}^{(1)}$ for other people than person 1 is built up from y_{2k}, y_{3k} … y_{nk} , which in turn are correlated with y_{1k} through the respective averages $y_{\cdot k}^{(2)}$, $y_{\cdot k}^{(3)}$ …. $y_{\cdot k}^{(n)}$, and thus its error term e_{1k}. However, the correlation between $y_{\cdot k}^{(i)}$ and e_{ik} is weaker than that between $y_{\cdot k}$ and e_{ik}. In accordance with this, simulation experiments (performed as above, but with somewhat more complex mathematical expressions) show that the estimate of the effect of $y_{\cdot k}^{(i)}$ is less biased than that of $y_{\cdot k}$: When a_1 once again is set to 0.26, $\hat{a}_1$ becomes 0.45. Also the effects of the other variables are closer to the true ones than in the simulation experiment reported above.

3.4. How Can the "Imitation Effect" be Estimated?

One will, in principle, get a better estimate of a_1 in model (7), or a similar model where there is at least one individual "identifying" variable whose community-level counterpart is excluded, by including more variables and thus picking up more of the unexplained variation. However, it is difficult in practice to be sure that the error term is

sufficiently small. As an illustration, I have estimated an extended version of model (1) from real data with more than 20 additional individual and community variables that have significant effects, and still got an estimate close to 1 for a_1 (0.94, as opposed to 0.99 with model (1)).

Obviously one should rearrange and estimate an equation such as (8) and use the coefficients $a_1a_4/(1-a_1)$ and a_4 to estimate a_1.

Also this approach hinges, of course, on the assumption that one particular individual variable z influences the dependent variable in focus, whereas the corresponding community variable is unimportant. In this particular example, z is the woman's age, and it seems indeed quite reasonable to assume that the corresponding community variable has no direct effect. Under this assumption, an effect of average age in a model where the average outcome is left out would reflect the importance of some unobserved community factors with which it is linked or that women's fertility is influenced by the number of children born to other women in the community, regardless of these women's age, and even though their number of children is partly a result of their age. The latter might well seem somewhat implausible, however, and suggest that the underlying idea that the behaviour of all uneducated women in the community is influential, without any further qualifications, be called into question. This complex issue of socio-demographic restrictions of the reference group is dealt with by Manski (1993).

Using this estimation technique on simulated data, which satisfy these assumptions that there would otherwise be much doubt about, gives a $\hat{a}_1$ estimate of 0.234, which is close to the true value.

3.5. Logistic Models

Let us now consider the following logistic model:

$$(12) \quad \log(p_{ik}/(1-p_{ik})) = a_0 + a_1y_{\cdot k} + a_2x_{ik} + a_3 x_{\cdot k} + a_4z_{ik} + a_5 q_{\cdot k}$$

where p_{ik} is a woman's probability of having at least three children (y=1), rather than fewer (y=0), and $y_{\cdot k}$ is the proportion of women in the cluster who have at least three children.

As above, realistic effect parameters are found by first estimating from real data, using a model that includes the proportion with three or more children among women with more than four years of education rather than that among those with less education.

Unfortunately, there is no simple expression such as (3) that can be used to calculate an average $y_{\cdot k}$. I use an iterative procedure instead. In the first step of the iteration, the observed $y_{\cdot k}$ is used to calculate p_{ik} from (12), and if a number drawn randomly from a uniform distribution is lower than this probability, y is set to 1 for that woman and is otherwise 0. An average $y_{\cdot k}$ is then calculated from the simulated y_{ik} and fed into (12) as a start of the second iterative step. The process converges after a few steps, in the sense that the average of the simulated y_{ik} equals the $y_{\cdot k}$ used in that step of the simulation.

On the basis of these outcomes that satisfy equation (12), with parameters

a_0=-0.490, a_1 = 0.710, a_2= -0.010, a_3=-0.320, a_4 = 0.050, and a_5= -0.070 ,

a logistic model is estimated. The estimates are

$\hat{a}_0$=-6.36, $\hat{a}_1$ = 7.41, $\hat{a}_2$= -0.003, $\hat{a}_3$=0.040, $\hat{a}_4$ = 0.054, and $\hat{a}_5$= 0.061,

which are vastly different. Most importantly, $\hat{a}_1$ is 10 times larger than a_1. Besides, the negative a_3 effect has disappeared, and the a_5 effect has changed sign. (Estimation based on real data gives a quite similar $\hat{a}_1$ (6.87), also when a large number of other variables are included.) (Note 4)

If I instead include the proportion among all women in the cluster who have at least three children, regardless of their education, the $\hat{a}_1$ estimate is slightly less different from the true value than in the simulation experiment above.

In another experiment, I group this overall proportion with three or more children in the cluster into four categories: 0-0.70, 0.70-0.81, 0.81-0.90, and 0.90-1.00. The women are quite evenly distributed across these four categories. After having first estimated a model that includes the categorized proportion among the better educated, I set the simulation parameters corresponding to the four categories to

a_{11} = 0 (reference group), a_{12}= 0.140 , a_{13} = 0.380, and a_{14} = 0.520.

The model is otherwise as (12). Estimation based on the simulated outcomes gives

$\hat{a}_{11}$ = 0 (reference group), $\hat{a}_{12}$= 0.633, $\hat{a}_{13}$ = 1.201, and $\hat{a}_{14}$ = 2.450.

This is the same type of model as estimated by McNay et al. (2003), except that they used the district (of which there are about 450 in India) as the level of aggregation, which gives a smaller bias, according to some additional model runs (details not

shown). Besides, their model was for contraceptive use, and a larger number of independent variables were included.

4. Conclusion

There are good reasons to believe that a person's demographic behaviour is influenced by that of other people in the community. However, one should not try to estimate that effect by simply including on the right-hand side of the model a measure of the average behaviour, constructed by aggregation within the sample. These simulation experiments, based on realistic data and parameters, have shown that the estimated effects of such variables can be very different from the true effects. Also the estimates of the other community variables can be strongly biased. It is possible to estimate an effect of the average of the dependent variable more indirectly, but only under very special assumptions that would be hard to defend.

As mentioned in the introduction, a different approach would be to use community data from other sources, for example a census or another survey. One might even consider splitting the sample into two, and construct a measure of the community average of the outcome variable from one part, while using the other to estimate the multilevel model including that variable. Such a community variable would not be a sum of the outcomes in the sample, and thus not so intrinsically linked with the error terms as the average-outcome variables considered above. However, there are nevertheless problems. All other community variables, for example average education, may be linked with unobserved factors that are important for the individual outcome, which would bias their effects, and this is obviously also the case for the average of the dependent variable. In fact, it is particularly hard to believe that the unobserved community factors associated with, say, the average behaviour in the community, as measured in other data, do not also have a bearing on the individual behaviour. While it would help to include many important variables in the model to pick up as much of the variation as possible, one should clearly be very careful to draw conclusions even when such "external" average-behaviour variables are used (Note 5).

5. Acknowledgements

Most of this work was done while the author was a visiting scholar at the East-West Center, Honolulu, with financial support from the Norwegian Research Council. The very helpful comments from three referees are greatly appreciated.

Notes

1. The classic linear regression equation can be written on matrix form as

   $\mathbf{y} = \mathbf{ma} + \mathbf{e}$

   where **y** is a column vector with elements y_{ik} for each of the N individuals in the sample and **e** is a column vector with error terms. **m** is a matrix with one row for each individual and the rows consisting of the independent variables, in this case the constant term 1, $y_{.k}$ and x_{ik}. The effect **a** is a column vector which in this case (model 5) includes the elements a_0, a_1 and a_2.
   As can be found in any textbook in econometrics, the OLS estimator is

   $\hat{\mathbf{a}} = (\mathbf{m}^t\mathbf{m})^{-1}\mathbf{m}^t\mathbf{y}$.

   Because simple matrix operations on the equation for **y** gives

   $\mathbf{a} = (\mathbf{m}^t\mathbf{m})^{-1}\mathbf{m}^t y - (\mathbf{m}^t\mathbf{m})^{-1}\mathbf{m}^t\mathbf{e}$,

   $\hat{\mathbf{a}}$ is $(\mathbf{m}^t\mathbf{m})^{-1}\mathbf{m}^t\mathbf{e}$ higher than the true **a**.
   The product $(\mathbf{m}^t\mathbf{m})$ has the 9 elements

N	$\Sigma y_{.k}$	Σx_{ik}
$\Sigma y_{.k}$	$\Sigma y_{.k}\, y_{.k}$	$\Sigma y_{.k}\, x_{ik}$
Σx_{ik}	$\Sigma y_{.k}\, x_{ik}$	$\Sigma x_{ik}\, x_{ik}$

   where the summation is over all individuals in the sample. $\mathbf{m}^t\mathbf{e}$ is a column vector with elements Σe_{ik} , $\Sigma y_{.k}\, e_{ik}$ and $\Sigma x_{ik}\, e_{ik}$. Let us assume that Σe_{ik} and $\Sigma x_{ik}\, e_{ik}$ are 0 (in this large sample). Thus, the difference between the estimated and the true effect of a_1 (second element of **a**) is simply the product between the second element in the second row of $(\mathbf{m}^t\mathbf{m})^{-1}$ and $\Sigma y_{.k}\, e_{ik}$. The former is $(N \Sigma x_{ik}\, x_{ik} - \Sigma x_{ik} \Sigma x_{ik})$ /d, where

   $d = N(\Sigma y_{.k} y_{.k} \Sigma x_{ik} x_{ik} - \Sigma y_{.k} x_{ik} \Sigma y_{.k} x_{ik}) - \Sigma y_{.k}(\Sigma y_{.k} \Sigma x_{ik}\, x_{ik} - \Sigma x_{ik} \Sigma y_{.k} x_{ik}) +$
   $\Sigma x_{ik}(\Sigma y_{.k} \Sigma y_{.k} x_{ik} - \Sigma x_{ik} \Sigma y_{.k} y_{.k})$.

   After some manipulation we get

   $\hat{a}_1 - a_1 = (N\Sigma y_{.k}\, e_{ik} - \Sigma y_{.k} \Sigma e_{ik})/((1-s)(N\Sigma y_{.k} y_{.k} - \Sigma y_{.k}\ \Sigma y_{.k}))$

   where

   $s = (N\Sigma y_{.k} x_{ik} - \Sigma x_{ik} \Sigma x_{ik})^2 / ((N\Sigma y_{.k} y_{.ik} - \Sigma y_{.k} \Sigma y_{.k})(\ N\Sigma x_{ik} x_{ik} - \Sigma x_{ik} \Sigma x_{ik}))$.

Recognizing the expressions for correlations and standard deviations, the formula can be written in a more compact form as:

$\hat{a}_1 - a_1 = \text{Corr}\,(y_{\cdot k}, e_{ik})\ \text{Std}\ e_{ik} / (\text{Std}\ y_{\cdot k}\ (1-\text{Corr}^2(y_{\cdot k}, x_{ik}))$

2. Ideally, one should split the error term into an individual-level (e_{ik}) and a community-level (u_k) contribution. When I use real data to estimate model (7) in Proc Mixed, except that I include a community-level error term and the average outcome among the better educated rather than that among the less educated, the standard deviation of the two error terms are 1.921 and 0.723, respectively. The standard deviation of the sum (v_{ik}) of the two error terms is, of course, 2.056 (=sqrt(0.723^2+1.921^2)) in accordance with the estimate reported above for a model with only an individual-level error term.
   When the number of individuals in each cluster is as small as here, the values of $e_{\cdot k}$ are markedly different from 0 and contribute substantially to the variation between clusters. Std $v_{\cdot k}$ is 0.966, which is higher than the 0.723 stemming from the u_k term. Std(v_{ik}-$v_{\cdot k}$), which is equal to Std(e_{ik}-$e_{\cdot k}$), is a measure of the within-cluster variation, and it is 1.811. In comparison, a model with only an individual-level error term with standard deviation 2.056 (which does not fit the data quite as well) gives between- and within-cluster variances of 0.681^2 and 1.938^2, respectively.
   When I include the average among the better educated rather than that among the less educated in the models and estimate from simulated data, the standard deviations of 0.723 and 1.921 are nicely replicated (0.731 and 1.918, respectively). However, when I include the average among the less educated in the simulation and estimation, the standard deviation of the community-level error term is estimated to be exactly 0, and that for the individual-level error term is estimated to be 1.819. This is very close to the true within-cluster variance of 1.811^2 in this sample. Similarly, when I use OLS for this estimation, the variance of the error term is estimated to be 1.819^2, and when I include only an individual-level error term with variance 2.056^2 in the simulation, the estimated variance is 1.946^2, which is close to the true within-cluster variance with that specification. In other words, all variance between communities disappears when these models that include the average of the dependent variable are estimated.

3. In principle, a more relevant "imitation" variable would be the average $Y'_k{}^{(i)}$ among all other women in the census-enumeration area (or a relevant sub-sample, such as that consisting of the less educated), given by

$$Y'_k^{(i)} = (Y_k N_k - y_{ik})/(N_k-1)$$

where Y_k is the census-enumeration area average and N_k the population size in this area, which is typically about 100 in these data. Because Y_k and N_k are not known, an alternative would be to use

$$y'_{\cdot k}^{(i)} = (y_{\cdot k} N - y_{ik})/(N-1)$$

with N taken as 100 in lack of more precise information. This is virtually the same as $y_{\cdot k}$, of course. Using

$$y_{\cdot k}^{(i)} = (y_{\cdot k} n_k - y_{ik})/(n_k-1)$$

instead is essentially to ignore that the woman in focus also "represents" about 10 others, and consider only the other women in the cluster as representative of other women in the area.

4. As above, the bias is less pronounced when I omit the person in focus before computing an average. With a simulation parameter a_1 that is set to 0.71 in (12), as above, the corresponding estimate is 1.36.
   A special case is that the cluster consists of only two persons, i.e. that

   $$\log(p_{1k}/(1-p_{1k})) = a_0 + a_1 y_{2k} + a_2 x_{2k} + a_3 x_{\cdot k} + a_4 z_{2k} + a_5 q_{\cdot k} \quad \text{and}$$
   $$\log(p_{2k}/(1-p_{2k})) = a_0 + a_1 y_{1k} + a_2 x_{1k} + a_3 x_{\cdot k} + a_4 z_{1k} + a_5 q_{\cdot k}$$

   A simulation experiment based on a smaller manipulated data set that includes clusters with only two women with little education yields a $\hat{a}_1$ effect of 1.51.
   The latter model bears some resemblance with that of McNay et al. (2003), which also included the contraceptive use of another woman in the household. They found that this variable had an extremely strong effect (that wiped out all household-level variation), which may well be true, but there is obviously good reason to doubt whether their estimate is reliable.

5. For such reasons, the bias in the models that include a combined average for women with little and those with more education is likely to be larger than suggested above, where it was assumed for simplicity that the average among the better educated was uncorrelated with the error term.

References

Blalock, H.M. (1984). "Contextual-effects models: Theoretical and methodological issues." *Annual Review of Sociology* 10: 353-372.

Erbring, L. and A. Young. (1979). "Individuals and social structure: Contextual effects as endogenous feedback." *Sociological Methods and Research* 7: 396-430.

Goldstein, H. (1995). *Multilevel Statistical Models, 2. edition.* Arnold: London.

International Institute for Population Sciences (IIPS) and ORC Macro. (2000). *National Family Health Survey (NFHS-2), 1998-1999: India*. Mumbai:IIPS.

Kravdal, Ø. (2002). "Education and fertility in sub-Saharan Africa: Individual and community effects." *Demography* 39: 233-250.

Manski, C.F. (1993). "Identification of endogenous effects: The reflection problem." *The Review of Economic Studies* 60: 531-542.

Moursund A. and Kravdal, Ø. (2003). "Individual and community effects of women's education and autonomy on contraceptive use in India." Forthcoming in *Population Studies.*

McNay, K., P. Arokiasamy and R.H. Cassen. (2003). "Why are uneducated women in India using contraception? A multilevel analysis." *Population Studies* 57: 21-40.