Multilevel analysis in health services research: a tutorial

Dag HOFOSS, Marijke VEENSTRA and Unni KROGSTAD

Foundation for Health Services Research, Nordbyhagen, Norway

Summary. - This article describes a well developed, but not yet widespread, statistical method. It presents its distinguishing characteristic (and pre-empts the misunderstanding that all it is about, is including predictors from more than one level). It walks the reader through a simplified example: how can multilevel analysis be used to explain variances in patient satisfaction with the way hospital work is organized. It also discusses the importance of the disciplinary setting for the extra benefit provided by multilevel analysis, as compared to traditional contextual analysis.

Key words: multilevel analysis, hospital and health services research.

Riassunto (Analisi a più livelli nella ricerca sui servizi sanitari). - L'articolo descrive un metodo che è stato sviluppato molto bene, pur essendo ancora molto diffuso. Presenta le sue caratteristiche (e prevenendo così il malinteso che si tratti soltanto di includere delle previsioni sulla base di più di un livello). Il metodo viene spiegato al lettore con un esempio semplificato: in quale maniera l’analisi a più livelli potrà essere usata per spiegare le variazioni nella soddisfazione del paziente con il modo nel quale il servizio dell’ospedale è organizzato. L’articolo discute inoltre l’importanza delle premesse delle discipline per ottenere il maggiore vantaggio offerto dall’analisi a più livelli nei confronti dell’analisi contestuale di carattere tradizionale.

Parole chiave: analisi a più livelli, ricerca su ospedali e sul servizio sanitario.

What is multilevel analysis?
The aim of this article

This article aims to introduce multilevel analysis to those who do not already know what it is, and to make clear the distinction between multilevel analysis and contextual analysis. The methodology is among the most recent developments in multivariate analysis, and it is well established. Despite a number of excellent papers [1, 2], it is not yet widely used in health related research. The example provided here uses the perhaps most widespread multilevel analysis programme, the MIWin (Multilevel analysis for Windows).

Multilevel analysis may be described as a method for solving a basic problem in research on sample data, namely that sampling often results in artificially small variance in the data. Underestimated variance produces too small standard errors and too narrow confidence intervals, and this leads to type I-errors: some of the effects one considers significant should have been written off as products of chance.

The cause of the problem is that sampling in large surveys is often done in several steps. Instead of pulling one random sample of the national population, one often first draws a random sample of municipalities, and in the next step a random sample of persons from those municipalities. That saves time and money. A side effect, however, is that the sampled individuals may be more similar than members of one simple random sample. Multilevel analysis is a method for handling this problem to avoid inflated t-values. In this perspective, multilevel analysis may be described as “regression with correct standard errors”.

The basic arguments for multilevel analysis

More exhaustive accounts often present three reasons for multilevel analysis [3, 4]. First, it is risky to look at covariances between aggregated variables alone. Wealthy neighbourhoods may have more burglaries, yet rich people may not be more likely to break in and steal. To study the relationship of individual wealth and the propensity to steal, one must also include data on individuals into the explanatory model. Likewise, it is risky to look only at covariances of characteristics of individuals. Personal deprivation may predict burgling, but may be a more powerful predictor where
opportunities abound. Multilevel analysis is a method for linking explanatory variables from different levels without committing such elementary fallacies.

Second, one should chose multilevel analysis even if the effect of social context may be studied by other methods. It is possible to study system effects on the relationship of variables at the individual level by doing separate analyses for each level 2-unit (in each group, at each school, in every hospital or county etc.). Multilevel analysis, however, makes do with far fewer parameters, and is thus a more efficient method.

The third reason is the one we mentioned right at the start. Data are often hierarchically structured, and any such nesting is likely to produce similarities between observations. A particularly obvious case is pupils, which are organized in classes, which belong to schools which are located in municipalities (and so on), but hierarchically structured units of observations is a general problem. As most populations are clustered, even simple random samples may contain units that are not independent. Given that one rarely can dismiss the possibility that social units are hierarchically structured, any sample may have “too little” variance. Not even simple random samples ensure that population variance is correctly estimated. To the extent that nesting is endemic to social systems, simple random sampling and cluster sampling are not fundamentally different. Because social reality is hierarchically patterned, population variances are almost always underestimated from sample data.

The problem of artificially small variance in stepwise cluster samples may be attacked in several ways. At least three solutions have been put forward. One is one should never do stepwise cluster samples unless forced by lack of time or money. That solution, however, is unpractical, for time and money often do preclude simple random sampling. Another recommendation is that when lack of resources dictate stepwise cluster sampling, one should multiply sample variances by 1.6 [5]. The third solution is multilevel analysis, in which dependencies among observations may be explicitly modelled.

**An example: patient satisfaction with the way hospital work is organized**

Patient satisfaction or dissatisfaction with the way hospital work is organized may be explained by patient variables as well as by hospital variables. The patient’s evaluation of the organization may for instance reflect the patient’s age. A nearly universal finding of patient satisfaction research is younger patients are more critical [6]. Whether or not the patient was diagnosed may also be consequential: patients having problems for which the hospital was unable to find a name may be more liable to consider hospital work not well organized. Indicators of staff job satisfaction may also prove important: patients treated in wards where fewer nurses like their jobs may have a less positive impression of how work was organized.

These are predictors at two levels. The first-mentioned two (age and nodiag: whether or not the patient’s problem was diagnosed) describe the patients, the third (nurjobsat: the percentage of nurses satisfied with the job) describes the ward. We have data on both: 9725 patients’ rating of their hospitalization at 123 wards, including their evaluation of aspects of hospital work organization. We also have data on the job satisfaction of the 2500 nurses working at these wards.

The data file is organized according to the demands of MIWin. The data were punched into an SPSS-file. Each patient constitutes one record. The top rows of the data matrix contain the values of patients having been treated at ward 1, the next rows the patients at ward 2 and so on down. On the patient variables each patient has his/her own value: the punch code assigned to the response alternative they ticked in the questionnaire. On the ward variable every patient at the same ward has the same value: the percentage of the nurses at that ward who were satisfied with their job. For instance, each patient at ward number 44 has the same score on the variable nurjobsat, namely 72, because 72 per cent of the nurses at that ward reported high job satisfaction.

Many patients are treated at more than one ward during their hospitalisation, here, they are classified as belonging to the ward at which they spent most days. Their evaluation of the hospital work may well have been coloured more strongly by what happened at the other wards at which they were treated, but in this article we disregard that possibility. One may also ask whether nurse job satisfaction was measured properly, but this article only aims to describe the multilevel method and does not go into this discussion (please note, also, that some of the estimates in Table 6 are adjusted for didactical purposes).

A note on terminology

The words multilevel analysis strongly suggests that the explanatory model employed contains predictors from more than one level. That is also how the above example was described: patient evaluation of the organization of hospital work may covary with whether or not the patient had his/her problem diagnosed and with the degree to which the ward was staffed with nurses who liked their work. The name, however, leads astray. To do multilevel analysis, one does not have to include into one’s explanatory model higher level predictors. It is of course a good idea to use explanatory variables from more than one level. But that is not what turns an analysis into multilevel analysis.
It is, despite the name, not the inclusion of variables from more than one level that makes an analysis multilevel. Contextual analyses are not multilevel analyses. That may be terminologically confusing, but that is how it is, and it is important to be aware of it from the start. The essential characteristic of multilevel analysis is the estimation of the variance of the regression coefficients that are hypothesized to vary between wards, and contextual analysis does not do that. What’s special about multilevel analysis is that it shows how different are the regression coefficients (not necessarily all, but at least one) across the higher level units to which the observations belong.

One can therefore do multilevel analysis without including a single explanatory variable. That follows from the defining property of multilevel analysis: the decisive factor is one estimates the between-units variance of at least one regression coefficient. Even a model completely void of explanatory variables may be a multilevel model, for the intercept is a regression coefficient, too. That is most easily seen by those who write it as \( \beta_0 \) and not \( a \), but, regardless of notation, the constant of the regression equation belongs to the same family as the effect coefficients \( \beta_1, \beta_2 \ldots, \beta_n \).

The simplest multilevel model is therefore the random intercept model: the model that has no explanatory variables, but estimates the total variance of the dependent variable, and decomposes it into variance in and between groups, that is, into level 1- and level 2-variance [7]. In our case of patient evaluation of how well hospital work was organized, it means modelling patient satisfaction like this (without explanatory variables):

\[
y_{ij} = \beta_0 + u_{0j} + e_{ij}
\]

In this equation, \( \beta_0 \) is the average satisfaction score of all patients, \( u_{0j} \) is a random effect at ward level (the level 2-residuals; the difference between the overall average of the data and the average score of each ward) and \( e_{ij} \) is a random effect at individual level (the level 1-residuals; the difference between each patient’s satisfaction and the average satisfaction score of the ward in which the patient was treated).

By definition, therefore, multilevel analysis begins, in spite of its name, not by listing relevant predictors at more than one level, but by the simple idea that subgroups in the data may differ with regard to our dependent variable. Before one begins to speculate about the possible causes of the differences in \( y \), one constructs the basis for multilevel analysis by thinking that maybe the observed \( y \)-values differ between units at the next level. That idea implies that we accept the possibility that the variance in the individual \( y \)-values may reflect characteristics of the units one level up. It is, of course, an important step forward to identify those characteristics, as well as the relevant explanatory variables at the level of the individual, and include both into one’s final model. But technically speaking one can do multilevel analysis with level 1-variables alone, even completely without explanatory variables.

The simplest multilevel model: the random intercept model (\( y = \text{patsatorg}, \text{no predictors} \))

The simplest multilevel model is, as said above, this:

\[
y_{ij} = \beta_0 + u_{0j} + e_{ij}
\]

This “empty model” estimates the average satisfaction score of the entire data set (\( \beta_0 \)), and the variance of ward averages around it (\( u_{0j} \)), and the variance of the patient scores around the average scores of the wards they were treated in (\( e_{ij} \)), and a goodness-of-fit measure showing how well the model fits the observed data.

The estimates are shown in Table 1, which, like the rest of the tables below is shown in the shape of the MWin-output.

The top line (\( \text{patsatorg}_{ij} \sim N(\mathbf{X}B, \Omega) \)) reminds the analyst that the usual linear regression assumption apply: the dependent variable (patsatorg = patient satisfaction with the organization of hospital work) is normally distributed.

The bottom line shows the model's goodness of fit-value (-2LL); this tells how well the model fits the data, that is how well one can predict patient satisfaction with the organization of hospital work by a regression equation with the parameters of this model. Lower values of -2LL indicate the model reproduces the observations better; the perfect explanatory model, which predicts the satisfaction scores of each patient without error, has a -2LL-value of zero. The figure 87674.440 is far above zero, but is in itself of no particular significance. One already knows that one is far off the mark if one predicts that every patient gave the same answer, namely the average answer of all patients. Yet, the -2LL becomes very instructive as new models are tried out: the question is how much is it

<table>
<thead>
<tr>
<th>Table 1. - Parameter estimates of the empty (random intercept only) multilevel model</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{Patsatorg}_{ij} \sim N(\mathbf{X}B, \Omega) )</td>
</tr>
<tr>
<td>( \text{Patsatorg}<em>{ij} = \beta</em>{0j} + \epsilon_{ij} )</td>
</tr>
<tr>
<td>( \beta_{0j} \sim N(0, \Omega_0) ); ( \Omega_0 = [17.112(3.066)] )</td>
</tr>
<tr>
<td>( \epsilon_{ij} \sim N(0, \Omega_{ij}) ); ( \Omega_{ij} = [474.048(6.840)] )</td>
</tr>
</tbody>
</table>

-2*loglikelihood(GLS) = 87674.440 (9725 of 9725 cases in use)

Patsatorg: patient satisfaction with the organization of hospital work.
reduced from one model to the next. The change in -2LL is a test of whether or not the present model is significantly better than the last one. Under the null hypothesis that the additional parameters have the population value of zero, the change in -2LL which follows from the inclusion of these predictors into the explanatory model is approximately $\chi^2$-distributed, its number of degrees of freedom being equal to the difference in the number of parameters of the two models which are being compared. The inclusion into the model of an extra predictor which does not reduce -2LL by more than 3.84 is therefore no significant improvement.

At our 0-100 satisfaction scale, the patients score, on average, quite high. The estimate for $\beta_0$ was 69.739, with a standard error of 0.445. The output in Table 1 does not only show the average satisfaction score of all patients, but also how satisfaction varies between wards. In the language of multilevel analysis, the regression coefficient $\beta_{0ij}$ is defined as random. As our file contains the variable ward no., the variance between wards can be estimated. All patients at the same ward have the same score on that variable (all patients at ward no. 1 have the score 1, all at ward no. 2 have 2 and so on). Thus the program knows which patients belong to which ward, and can calculate the average score of the entire data set as well as for each ward, and also the variance between and within wards. The between-ward-variance was 17.112 (standard error 3.066). The data set’s total variance is the sum of the variances at level 1 and level 2: $474.048 + 17.112$, that is $491.160$. Relative to that, between-ward-variance is small. The fraction of the total variance that is produced by characteristics of the wards (the intra-class correlation coefficient) is only $17.112/491.160 = 3.5 \%$. It is, however, statistically significant ($17.112/3.066 = 5.58$).

Multilevel analysis is particularly useful when a considerable part of the variance in y is between units at higher levels, that is, when the intra-class correlation coefficient is high. When it is quite small, like here, characteristics of the units at the next level upwards (in our case the wards) do not affect the individual scores on the dependent variable much. Under such circumstances, the results that are produced by multilevel analysis do not differ much from those produced by ordinary regression. Still, multilevel analysis provides the additional information of which fraction of the data set’s total variance is at level 2.

**A multilevel model with one explanatory variable (at level 1: $x_1 =$ age)**

We include as our model’s first predictor the variable age. We ask ourselves whether it should be considered to have the same effect on patient satisfaction in all wards, and decide it should. We therefore define it as having a fixed effect. That implies that we ask MIWin to estimate only its average effect (and not also the effect coefficient’s variance between wards). The model’s parameter estimates are shown in Table 2.

-2LL is now 87390.650, that is as much as 283.790 below the -2LL-value of the random intercept-model with no explanatory variables. Relative to that model, one additional parameter was now estimated. The $\chi^2$-table says that such a large improvement in fit at the cost of only one degree of freedom is highly significant: $p < 0.001$.

The effect coefficient for age is interpreted as in ordinary linear regression. It shows that patients one year older on average scored 0.210 points higher on the (0-100) scale for satisfaction with the way hospital work was organized. The coefficient is highly significant: $0.210/0.012 = 17.5$.

To the extent that parts of the explanation of why patients regard the organization of hospital work differently is that patients of different ages are differently critical, the inclusion of age as a predictor should reduce the variance in satisfaction. And it did: the inclusion of the predictor age reduced the total variance in the data by 15.580. In the empty model it was 491.160, in Table 2 the residual variance is down to $14.661 + 460.920 = 475.581$.

### Table 2. - Parameter estimates from a model including the predictor age

<table>
<thead>
<tr>
<th>$Patsatorg_{ij}$</th>
<th>$\sim N(XB_{ij}, \Omega)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Patsatorg_{ij}$</td>
<td>$= [57.490(0.834) + u0j + e0ij]$</td>
</tr>
<tr>
<td>$u0j$</td>
<td>$\sim N(0, \Omega_u) : \Omega_u = [14.661(2.730)]$</td>
</tr>
<tr>
<td>$e0ij$</td>
<td>$\sim N(0, \Omega_e) : \Omega_e = [460.920(6.651)]$</td>
</tr>
</tbody>
</table>

-2$loglikelihood(GLS) = 87390.650 (9725 of 9725 cases in use)

Patsatorg: patient satisfaction with the organization of hospital work.

### Table 3. - Parameter estimates from a model including predictors age and not having been diagnosed during hospitalization

<table>
<thead>
<tr>
<th>$Patsatorg_{ij}$</th>
<th>$\sim N(XB_{ij}, \Omega)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Patsatorg_{ij}$</td>
<td>$= [58.128(0.846) + u0j + e0ij]$</td>
</tr>
<tr>
<td>$u0j$</td>
<td>$\sim N(0, \Omega_u) : \Omega_u = [14.403(2.689)]$</td>
</tr>
<tr>
<td>$e0ij$</td>
<td>$\sim N(0, \Omega_e) : \Omega_e = [460.131(6.639)]$</td>
</tr>
</tbody>
</table>

-2$loglikelihood(GLS) = 87372.620 (9725 of 9725 cases in use)

Patsatorg: patient satisfaction with the organization of hospital work.
Table 4. - Parameter estimates from model including predictors age, not having been diagnosed and the percentage of nursing staff reporting high job satisfaction

\[
P_{\text{atsatorgi}} = N(X'\beta, \Omega)
\]

\[
P_{\text{atsatorgi}} = [\beta_0 + 0.206 \times \text{age} + -2.919 \times \text{nodiag} + 0.098 \times \text{nurjobsat}]_i
\]

\[
\beta_0 = [50.681(2.201) + u_0j + e_{0ij}]
\]

\[
\beta_j = N(0, \Omega_j) : \Omega_j = [12.072(2.391)]
\]

\[
\beta_k = N(0, \Omega_k) : \Omega_k = [460.217(6.640)]
\]

\[
-2 \times \text{loglikelihood(IGLS)} = 87360.220 (9725 of 9725 cases in use)
\]

Patsatorg: patient satisfaction with the organization of hospital work.

One more explanatory variable at level 1: whether or not the patient’s problem was diagnosed (\( \chi^2 = \text{nodiag} \)).

It may well be imagined that patients with health problems that the hospital is unable to diagnose, may be more liable to consider hospital work not well organized.

We define this predictor’s effect as fixed, too, that is, we assume that not being diagnosed has the same effect on patient (dis-)satisfaction in all wards. The parameter estimates are shown in Table 3.

The model’s -2LL-value is now 87372.620. The inclusion of the predictor nodiag improved its fit by 87390.650 - 87372.620 = 18.030. The improvement only cost one degree of freedom, and is thus highly significant (\( p < 0.001 \)).

The effect of age was not much changed: its coefficient is now 0.205. Its standard error did not change, and age is still a highly significant predictor (0.205/0.012 = 17.08).

Patients who did not have their problem diagnosed during hospitalisation, judged, on average, the organization of hospital work 2.947 points lower. This coefficient, too, is significantly different from zero (2.947/0.694 = 4.25).

The difference between the results from multilevel and ordinary regression

The regression model presented above - patient satisfaction with organization of hospital work = f(age and having had the problem diagnosed or not and having been treated in a ward in which nurses were satisfied with their job) - might have been used in ordinary regression analysis as well. The effects of the patient’s age, of not getting a specified diagnosis and of being treated in a ward with a specified percentage of nurses being satisfied with their job might have been analysed without taking into account the nested structure of the data. To demonstrate the difference in results, we have performed an ordinary regression analysis (SPSS, version 11.5) of the same data set. The results are shown in Table 5.

The effect coefficients did not change much. Their standard errors, however, as signalled in this article’s first paragraph, increased - reducing t-values and
thereby the risk of committing type 1-errors, that is, believing in effects that might have been due to chance. In this example, the changes were relatively small. In other cases, the differences may be crucial.

**Random effects: regression coefficients that vary across wards**

So far we have assumed that the effect of each predictor on patient evaluation of the organization of hospital work was the same at each ward: all predictors have been defined as fixed. We have thus only developed the explanatory model’s fixed part, and ignored the possibility that effects can vary across wards. The idea that effects may vary across units at higher levels is, however, one of the advantages of multilevel analysis.

One may for instance suspect that the effect on patient satisfaction with the organization of hospital work of not having had one’s health problem diagnosed may vary from one ward to another. At one ward, staff may take care to explain the diagnostic problems, at another the patient may be left to ponder by himself the question of what went wrong – and in the latter case the patient may more easily conclude that hospital work was not well organized. To see whether the effect of the predictor nodiag varies across wards, we redefine this predictor to be random, that is, we remove the very restrictive estimation assumption that its effect is exactly the same for every ward, and instead ask for the between-wards variance of these effect coefficients: how different are they. MlWin then also estimates the covariances of the intercept and each random predictor’s effect, as well as the covariances between the effects which have been declared random.

Our model has only two levels: patient and ward. The effect of any patient variable (level 1-variable) can be defined as random. The predictor nurjobsat, on the other hand, is at the model’s highest level. Our model knows no higher-order units between which its effect can vary. Now, there is nothing about the variable itself that prevents its effect to be regarded as random. Its fixedness follows solely from the fact that our model does not specify any level above the wards. It might well have done so, for wards belong to departments, departments to hospitals, and hospitals to health regions (to mention three additional levels). Had our data file included a variable which said which department each ward belonged to, we might have defined the effect of this - and any other - ward level predictor as random, and had MlWin estimate the variance of its effect across departments. In this article, however, we restrict ourselves to two levels.

At this point, a note should be made about the subscripts that have accompanied the variables in the above figures: nurjobsat has the subscript j, whereas age and nodiag are subscripted i. Any multilevel variable having a subscript including the letter i is a level 1-variable. Ageij means age is a level 1-variable: each patient (i), at each ward (j) has an age-value. A variable which is subscripted j only, is at level 2. Nurjobsatj means this variable is at ward level: the percentage of nurses reporting high job satisfaction is a ward characteristic.

Models with three levels group their predictors by level by three subscript combinations: ijk, jk and k. A variable subscripted ijk is at level 1, and describes, e.g., patients. The subscript jk shows the variable is at level 2, describing, e.g., wards. Any variable subscripted k is at level 3, describing, e.g., departments.

Individuals are normally clustered in more than one type of units at higher levels, and the clusters may overlap. Hospital patients belong to wards, but they also belong to e.g. diagnostic groups and municipalities. Multilevel analysis can handle such criss-crossings, but that subject falls outside the scope of this article.

A predictor at level 1 is a predictor whose effect might vary between units at higher levels. But it doesn’t have to do that. The fact that its effect might vary between (e.g.) wards, does not compel us to believe that it actually does, and ask for its variance to be estimated. The fact that a variable is at level 1, and in our two-level model has the subscript i, does not demand that its effect be defined as random. We can, on theoretical or empirical grounds, freely decide to believe that that variable’s effect is the same at all wards or that it varies between wards. In the first case we define its effect as fixed, and request that the programme estimates only its average effect (its effect in the entire data set, seen as one group). In the second

<table>
<thead>
<tr>
<th>Table 5. - Ordinary linear regression (SPSS) and multilevel regression (MlWin) results: coefficients, standard errors and t-values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td><strong>SPSS</strong></td>
</tr>
<tr>
<td>β</td>
</tr>
<tr>
<td>s.e.</td>
</tr>
<tr>
<td>t-value</td>
</tr>
<tr>
<td><strong>MlWin</strong></td>
</tr>
<tr>
<td>β</td>
</tr>
<tr>
<td>s.e.</td>
</tr>
<tr>
<td>t-value</td>
</tr>
</tbody>
</table>

Cons: constant.
Nodiag: not having had the health problem diagnosed.
Nurjobsat: percentage of nurses reporting high job satisfaction.
s.e : standard error.
Multilevel analysis in health services research: a tutorial

Case we define it as random, and ask the programme to estimate both its average effect and the variance of the effect coefficients at each ward.

Whether a variable has been defined as fixed or random is seen from its effect coefficient's subscript, and not the variable's. If the effect of a predictor $x_4$ is defined as random (varying across wards), it shows up in the output as $\beta_{4j}$.

Redefining the predictor nodiag from fixed to random

The model described above has two levels, and thus no higher-level units between which the effect of our level 2-predictor nurse job satisfaction can vary. That predictor, therefore, can only belong to the model's fixed part. But the effects of the two level 1-predictors age and not having had one's problem diagnosed may vary across wards. For the sake of this article we chose to believe that while the effect of age is the same at all wards, the effect of the predictor nodiag varies across wards (in a real analysis we would have had to say why, here we just postulate it).

The results from the model that regards the effect of not having had a diagnosis as varying across wards is shown in Table 6.

The elimination of the restriction that the effect of not having been diagnosed must be the same at each ward, produced a model with a better fit. $\text{-2LL}$ was reduced by $87360.220 - 87352.100 = 8.12$. As can be seen from the output, not just one, but two, additional parameters have been estimated. One is the between ward-variance in the regression coefficient that expresses the effect of not having been diagnosed during hospitalization, the other is the covariance of the regression equation's intercept and the varying effect coefficient. The improvement in fit was won at the cost of two degrees of freedom. At two degrees of freedom, a -2LL reduction of 8.12 is statistically significant ($p < 0.01$).

The two effects that were supposed not to vary by ward, changed very little. Patients one year older are now 0.205 points more satisfied with the way hospital work was organized, and patients treated at wards where 1% more of the nurses reported high job satisfaction still score 0.094 points higher. Both coefficients are still significant.

Table 6 shows, like Table 4, the average effect of not having been diagnosed. It is now estimated to be -2.982, and it still looks significant: $-2.982/0.805 = -3.70$. The u-matrix, which in the tables above just contained the between-wards variance of the regression equation intercept, in Table 6 turns out to be a variance-covariance matrix for all coefficients in the model’s random part. In our case, it shows the between-wards variances of the intercept and not having been diagnosed, as well as the covariance of the ward intercepts and random effect coefficients. If the model included more than one random effect predictor, the u-matrix would also have shown the between-wards variance(s) of that (those) variable(s), its (their) covariance(s) with the ward intercepts and the covariances of all the random effect coefficients.

Repeating in Table 7 the variance-covariance matrix of Table 6.

\[
\begin{align*}
\beta_0 & \sim N(\mathbf{X}\mathbf{B}, \Omega) \\
\beta_0 & = 63.266(2.055) + u_0 + e_0 \\
\beta_2 & = -2.982(0.805) + u_0 + e_0 \\
\begin{bmatrix} u_0 \\ u_2 \end{bmatrix} & \sim N(0, \Omega_u) : \Omega_u = [8.519(2.398), -5.106(2.431)12.650(4.269)] \\
\begin{bmatrix} e_0 \\ e_2 \end{bmatrix} & \sim N(0, \Omega_e) : \Omega_e = [459.013(6.651)] \\
\text{Patsatorgij} & \sim N(\mathbf{X}\mathbf{B}, \Omega) \\
\text{Patsatorgij} & = \beta_0\text{cons} + 0.205(0.012)\text{age} + \beta_2\text{nodiag} + 0.094(0.027)\text{nurjobsat} \\
\text{Patsatorgij} & = 63.266(2.055) + u_0 + e_0 \\
\text{Patsatorgij} & = \beta_0(\text{cons}) + 0.205(0.012)\text{age} + \beta_2\text{nodiag} + 0.094(0.027)\text{nurjobsat} \\
\end{align*}
\]

Table 7.

<table>
<thead>
<tr>
<th>$\beta_0$ (intercept)</th>
<th>$\beta_2$ (effect of $x_2 = \text{nodiag}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.519 (2.398)</td>
<td>-5.106 (2.431) 2.650 (4.269)</td>
</tr>
</tbody>
</table>

Nodiag: not having had the health problem diagnosed.
satisfaction between patients who have been diagnosed and patients who have not, was bigger where the ward intercept was lower (where patients on average were less satisfied with the organization of hospital work). The relatively large standard error (8.269) of the covariance indicates, however, that it is not significant.

The upper figure in Table 7’s leftmost column, $u_{0j}$ (8.519) is the between-ward variance of the regression equation intercept. Together with the figure in the $c_{ij}$-matrix (459.013) it adds up to the unexplained variance of patient satisfaction with the organization of hospital work: 8.519 + 459.013 = 467.532. In the empty model (Table 1), the total variance in $y$ was 17.113 + 474.048 = 491.161.

Explained variance

The multilevel analysis has explained relatively more of the $y$-variance at ward level (which was reduced from 17.1 to 8.5) than at patient level (where the reduction was from 474.0 till 459.0). The amount of explained variance reflects, like the $R^2$ of ordinary regression analysis, the proportional reduction in prediction error. The estimations are, however, slightly more complicated. First, there are in multilevel analyses as many $R^2$s as there are levels in the model - in our two-level model an $R^1_2$ and an $R^2_2$. Secondly, the calculation of the $R^2_2$ requires the introduction of a factor that adjusts for group size. Thirdly, random effects models will, on principle, need special formulae [8] - in practice, however, a simpler technique is used.

As the estimated values of $R^1_2$ and $R^2_2$ usually change very little when random regression coefficients are introduced, one may re-estimate the model as a random intercept model where all explanatory variables are defined as fixed. This will usually yield values that are very close to the values that might have been calculated by the special formulae [7].

In our example, then, we will use the estimates from Table 4 to calculate the $R^2$s. The $R^1_2$ (the proportional reduction in prediction error at patient level) is estimated “in the usual way”, as the difference between 1 and the ratio of the residual variance in the dependent variable (which in Table 4 was 12.072 + 460.217 = 472.289) and its original total variance (in the empty model of Table 1, which was 17.112 + 474.048 = 491.160). The $R^1_2$, then, is $1 - (472.289/491.160) = 0.038$.

To calculate the ward level $R^2_2$, one must first calculate the residual total variance at level 2 as the sum of the residual level 2-variance and the residual variances at level 1 adjusted for group size. In school research, where groups (classes) are often quite similar in size, the latter operation may just mean dividing the level 1 residual variance by the usual group size (say, 25 - if that is the typical class size in the data set). In hospital research, where units may differ considerably in size (our 123 wards differed from 12 to 288 patients treated in the period studied), one may use the harmonic mean, defined as the number of wards divided by $\sum (1/n_j)$, the sum of the inverse of the ward sizes. In our data set that sum is $1/12 + 1/16 + 1/38 + \ldots + 1/288 = 3.324$, producing an adjustment factor (a mean ward size) of $123/3.324 = 37.004$. Using the estimates of Tables 1 and 4, $R^2_2$ – the proportional reduction in error variance at level 2 – is estimated to 18.1%: $1 - (472.289/37.004 + 12.072)/(474.048/37.004 + 17.113) = 0.181$.

The implications for the utility of multilevel analysis of the distribution of variance by levels

A very important result was produced already by the empty model of Table 1: the dominating part of the variance in patient satisfaction with the organization of hospital work is variance between patients and not between wards. That may be seen as a major health services success. An important health care goal is that it should not make a difference whether a patient is sent to ward A or ward B. Hospitals may take special care to counteract unintended differences between wards, and these data may indicate that they succeed.

A number of factors related to the typical hospitalization for acute somatic illness may also reduce between-ward variation. It is easy to see that the effect of the hierarchical structure in data on pupils and schools may be strong. Classmates often grew up in the same neighbourhood, with not very different social backgrounds (as opposed to pupils in other classes in other schools), they have the same teacher(s) and go through the same curriculum in the same way (as opposed to pupils in other classes in other schools) and - not leastly important - they constitute a social unit for years and influence each others' building of personal identity both through direct communication and by building a group culture and developing norms together (in other ways than in other classes in other schools).

In health services research, level 2-variance is probably often smaller. The patients at a ward are not recruited from a particular social unit, rarely did they grow up together, and they do not often have the same socio-demographic characteristics. Patients with related diseases go to the same ward and are treated by the same personnel and the same methods. But they would have had approximately the same treatment at another ward. Furthermore, with an average length of stay of less than five days (of which the patients, and/or their ward neighbours, at least part of the time are so sick that they mostly keep, or are kept, by themselves) they do not have much time to develop a social system that powerfully influences their reactions and evaluations.
For this reason, it is hardly a coincidence that multilevel analysis is rooted in school research. Most introductions to multilevel analysis take their examples from schools - and the programme MIWin is developed in the large “multilevel models project” at the University of London’s Institute of Education. Multilevel analysis is probably particularly well suited to the study of systems with a marked socializing effect, and is not designed with patient experiences research in mind.

Yet, even if the need for and usefulness of multilevel analysis vary from one research field to another, we are not arguing that it is not useful in health-related research. On the contrary: this article is meant to promote interest in multilevel analysis in health services research, even in patient satisfaction research. A fair amount of medical and sociological research indicate that one should not disregard the community of patients at wards. As has been pointed out [9, 10], patient expectations are not just fixed ideas that patients bring into the hospital to measure their experiences by. Instead, they should be seen as part of a culture developing in the course of the patients’ interaction with staff and fellow patients [11]. Also, medical treatment is so far from being standardized that it may make a difference where the patient is treated. It has been shown for decades that medical treatment of the same problem varies considerably even within limited geographical areas, maybe even from doctor to doctor [12-15]. Therefore, although the exact strength of the effects of the hierarchical structure in data on hospital patients may be an empirical question, one should not disregard the possibility that patient satisfaction research and other hospital sociology may need to identify the effects on clustering of observations. Also, health services research is not only about patient experiences, it is also, among other things, about staff job satisfaction, where intra-class correlations may be much stronger.

Multilevel analysis software

The multilevel analysis presented above was done by MIWin. The data set was originally in an SPSS file. In our case the file was a merge of two SPSS data files, one containing patient data (of which this example used the variables age and not having been diagnosed during hospitalization) and another containing ward data (of which we used percentage of nurses reporting job satisfaction). The merged SPSS file was supplied with variables that identified (by a rank number) the wards as well as the patients, and sorted by ward rank and patient rank to make record no. 1 patient 1 in ward 1, record no. 107 the 107th and last patient in ward 1, record no. 108 patient 1 in ward 2 and so on. The sorted data file was transformed into an ASCII text file and read into MIWin as an MIWin worksheet to produce the output shown in the above tables.

MIWin may be the most widely used multilevel analysis program, but it is not the only one. Multilevel analysis can be done by other programs specially developed for multilevel modelling and by multilevel modules in general purpose program packages. Perhaps the best known other special purpose programs are HML [16] and VARCL [17]. Hox [18] gives detailed descriptions of multilevel analyses in HML and VARCL.

Of the general purpose program packages, SAS [19] and Stata [20-22] have full-fledged possibilities for multilevel analysis. SPSS and BMDP have been slower to include such options but are gradually developing modules that allow multilevel modelling.

Web sites

A very useful web address is http://multilevel.ioe.ac.uk, at which the Centre for Multilevel Modelling in London provides updated information on MIWin and offers a number of links to other relevant web sites, including 18 multilevel software home pages, among them SAS and Stata, as well as a link to an e-mail multilevel modelling discussion list.

Received on 11 February 2003.
Accepted on 3 April 2003.

REFERENCES


