CHAPTER 4

4 REVIEW OF APPLICATIONS

In this Chapter we introduce a number of practical examples and applications of hierarchical data structures that were previously described within the thesis. These examples cover the whole application area of multilevel modeling, from educational data analysis to health analysis, from social statistics to survey research, as well as meta-analysis, repeated measures analysis and so on. It is profound that the applications reviewed here are only a minor sample of the thousands printed articles dealing with applications and data analysis by using, among others, multilevel techniques. However, the main reason they were chosen is that they have formed the basis for theoretical discussions as well for other applications in the same or other areas of interest, from the same or other authors.

In this Chapter we will focus more on the description of the example, the data description and analysis, the statistical methods and techniques and the discussion of the results rather than the results themselves. We will also mention other relative to the examples, reviews and articles from the same or other authors, which were based or "inspired" from the particular articles. We split our review into examples taken from the area of education and applications in all other areas of interest.

The scope of this Chapter is, therefore, to show if Multilevel Analysis is handly applicable in practice and if the theoretical advantages of the Multilevel Techniques, as discussed in the previous Chapter, are also present in practical situations derived from various research areas.

4.1 Applications in Education

Analysis of educational data, students and organizations performance is the area were multilevel model techniques were firstly introduced and have been used to a great extent, due to the profound hierarchical structure of the data measured in this area. We start our review from an introductory multilevel analysis of school examination results (Goldstein et al, 1993), which, however, form the basis to almost all the analysis in these areas.

'A Multilevel Analysis of School Examination Results'

a. INTRODUCTION

Goldstein et al. (1993) examined data on examination results from inner London schools in relation to intake achievement, pupil gender and school type. The examination achievement, averaged over subjects, was studied as well as achievement in the separate subjects of mathematics and English. Multilevel models were fitted, so that the variation between schools could be studied. Specifically he focused on two measures of intake achievement for each school in the study, and examined the interpretational issue by studying the dimensionality of school differences.

b. DATA

The data were examination results from 5748 students in 66 schools in six Inner London Education Authorities. These students had data on their General Certificate of Secondary Examination (GCSE) grades in Mathematics and English, together with a total score for all the subjects taken in that examination. For mathematics and English, a scale ranging from 0 (no grade awarded) to 7 (grade A) was used in the analysis and, for the total score, the scale ranged from 0 to 70. These students also had scores on a common reading test taken when they were 11 years old-the London Reading Test (LRT) (Levy & Goldstein, 1984) and were graded also into three categories on the basis of a verbal reasoning (VR) test at 11 years (Nuttall et al., 1989). All three scores were scaled to have mean zero and standard deviation 1. The pattern was similar for all three response variables.

The original number of students on whom some examination data had been obtained was 8857 in 74 schools. Students were omitted from the analysis if they did not have both intake measures. Where students did not take an examination they were given a score of 0, the same as if they obtained an ungraded result. The exclusion of these students resulted in a sample with a higher total examination score, 23.7 as opposed to 20.0. As the author pointed out this differential loss of students with lower examination achievements needs to be born in mind when interpreting the results, and is a persistent problem with data of this kind.

Two separate models had been fitted to the data. The first analyses the total examination score and the second is a bivariate analysis of the English and

mathematics scores. All the response variables had been transformed using normal scoring to conform as closely as possible to multivariate normality.

c. TOTAL EXAMINATION SCORE

The explanatory variables used in this analysis were as follows:

- standardized London reading test (LRT);
- verbal reasoning category;
- gender;
- school gender (mixed, girls, boys);
- school religious denomination (State, Church of England, Roman Catholic, other).

Formally, the model was written as follows:

$$y_{ij} = \sum_{h=0}^{5} \beta_h x_{hij} + \sum_{h=6}^{10} \beta_h x_{hj} + \sum_{h=0}^{2} u_{hj} x_{hij} + \sum_{h=0}^{1} e_{hij} x_{hij} \quad (4.1)$$

where i refers to student and j refers to school. Throughout this equation the subscript 0 refers to the constant term (= l), the subscript 1 to LRT and 2 to the dummy variable for VR group 1. Subscripts 3-5 refer to the square of LRT, the dummy variable for verbal reasoning group 2, and the dummy variable for gender. The subscripts 6-10 refer to the five school-level defined i.e. the 'fixed part'. These are:

- Girls-mixed school
- Boys-mixed school
- CE-State school
- RC-State school
- Other-State school

The first summation refers to the explanatory variables defined at the student level, the second to those defined at the school level, the third to the random part of the model defining variation at the school level, that is level 2, and the fourth summation defines the random variation at the student level, that is level 1. We also have, at level 2,

$$\text{var}(u_{hj}) = \sigma_h^2 \quad (4.2a)$$

$$\text{var}(e_{hij}) = \sigma_{eh}^h \quad (4.2b)$$

The level 1 contribution to the variance is:

$$\sigma_{e0}^2 + 2\sigma_{e01} + \sigma_{e1}^2 x^2 \quad (4.3.)$$

That is, a quadratic function of x and the individual level variances and covariance in this expression do not have separate interpretations. On the other hand, since the x_{hij} are defined at level 1, the level 2 variances and covariances are interpreted directly as between-school variances and covariances for the relevant coefficients. Several exploratory models were fitted and the main results for the model found to give the most satisfactory fit are the following:

- The level 1 (between students) variance is a linear function of LRT score given by: $Variance = 0.55 + 0.092LRT$
- Likelihood ratio test statistics for:
  - (a) Level 1, LRT (covariance): $X_1^2 = 66.0, P < 0.001$
  - (b) Level 2, VRl (VR2,VR3) variances and covariance: $X_3^2 = 11.0, P = 0.012$
  - (c) Level 2, LRT variance: $X_3^2 = 24.7, P < 0.001$.
- The effect of school gender is small and the differences are about the same order of magnitude as the estimated standard errors.
- There seems to be a small advantage for those attending Roman Catholic schools.
- Girls do better than boys
- There are large differences between those in the different verbal reasoning categories and there is a strong quadratic relationship with LRT.
- The relationship between examination score and LRT varies, as does the difference between verbal reasoning categories 1 and 3, with high positive correlations.
- At the student level the variance increases with increasing LRT score.

After having fitted the model it was possible to estimate the level-2 residuals. These are obtained by estimating the regression model with the (unknown) residuals as responses. The resulting estimates are often known as 'shrunken' estimates since, like all regression predictions they have smaller variances than that of the true values. To illustrate the implications of the model, the authors formed particular extreme combinations of the school residuals. So for each school they calculated, the two combinations $u_{0j} - 2u_{1j}$ and $u_{0j} + 2u_{1j} + u_{2j}$ that is, first the estimated school 'effect'

for a student with an LRT score of - 2, the approximate lower 2.5th percentile, and in verbal reasoning group 2 or 3, and second the estimated 'effect' for a student at the approximate upper 2.5th percentile and in verbal reasoning group 1.

It was concluded that there is a positive correlation between the school 'effects' for the low and high achievers on intake. Nevertheless, there were some schools with below average values for the low achievers which had above average values for the high achievers, and vice versa. This emphasizes the point that schools appear to be differentially effective for different kinds of students.

Then, approximate 95% confidence intervals for estimate of the intercept residual were constructed and plotted. That is the school 'effect' estimated at the mean LRT score for those in verbal reasoning groups 2 and 3. It was noted that these intervals were calculated separately for each residual, and were based upon the estimated standard error, which in general are an underestimate of the true standard error. For comparing any two particular schools, the usual significance test and confidence interval procedures were used. It was seen that there is a very considerable overlap of intervals, which suggests that it is not possible statistically to discriminate easily between schools. In particular, there are no natural division points in the sequence of estimates, which would allow the authors to classify schools into homogeneous subgroups.

d. JOINT ANALYSIS OF ENGLISH AND MATHEMATICS SCORE

The next step was the analysis of the English and mathematics examination scores. These were chosen because, in principle, these examinations are taken by all students. It would be possible to carry out a joint analysis of these two scores together with the total score on the other subjects, but for simplicity of interpretation the authors restricted to just the two. Also for simplicity, they used only the student level variables as explanatory variables, and at the between-school level they used only the intercept and LRT coefficient as random variables.

A multivariate model was fitted by treating the multiple variates within each student as the level 1 classification. In this case, therefore, there were two level 1 units within each student (level 2) with schools at level 3. It was concluded that:

- In the fixed part of this model the average difference between girls and boys is 0.1 units in favor of the boys for mathematics and 0.23 units in favor of the girls for English.

- For LRT and verbal reasoning categories, there is little difference between the relationships for maths and English.
- In the random part of the model the LRT coefficients for English and maths do not vary greatly. The standard deviation for maths is 0.006 units while that for English is only 0.003.
- While schools differ in terms of overall maths and English performance, at least for English, there is little differential effect according to LRT at intake. The intercepts for maths and English have a small correlation (0.09), and there is only a moderate correlation for the intercepts and LRT coefficients for both maths and English.
- At the student level the between-students variation decreases from VR1 to VR3 category students. This is similar to the finding in the analysis of total score, where the lower achieving intake students (based on LRT) had smaller variance
- If a model is fitted with just a variance term for mathematics and English and a covariance term at level 1, the effect is to increase the standard errors for both the fixed part of the model and the level 2 random parameters by up to 20%. The estimates of the coefficients and parameters themselves do not change appreciably, but the decrease in precision emphasizes the importance of accurate level 1 modeling.

To illustrate the relationship among the school level residuals or 'effects', the residual estimates for the two intercepts, that is at the mean LRT score were plotted. It was shown that there is little relationship between English and maths performance. The school with the greatest English residual was only average for maths and one of the schools with high maths residual had a low value for English. This relationship is for those students with average LRT scores. Since the LRT coefficients vary across schools, the relationship between maths and English residuals will also vary with the LRT score.

e. DISCUSSION

The analysis leaded to some important comments mentioned by the authors:

- There is an association between the between-student variance and the intake achievement score, with increasing variation as the intake achievement increases. This is of substantive interest and it is also important to incorporate

it in the model since it helps to ensure that the overall model is correctly specified and will generally improve the precision of the remaining parameters.

- The authors criticize on the reliability of the explanatory variables used in the analysis (LRT score or the VR band allocation). Also, whether the use of particular variables (verbal reasoning and reading achievement measures) are entirely satisfactory to adjust for intake. They comment that, ideally, when total examination score is used as the response, initial achievement measures should cover the full range of school subjects taken in the examination
- They also criticize 'league tables' (published Tables with the average General Certificate of Secondary Education examination results in England and Wales, used by parents in order to choose schools and colleges) whether or not these are adjusted for intake achievement and whether or not multilevel modeling has been used. The ordering of school effects depends on the intake achievements of students as well as the curriculum subject being examined. Also, a study of residuals differentiated by intake achievement and by subject, can suffice as a screening device and as feedback to individual schools about potential problems Even then however, the ability of comparison between schools based on residuals is still controversial. Fine distinctions and detailed rank orderings are statistically invalid.

Harvey Goldstein as the “leader” and a large group of other scientists have introduced the “Institute of Education” in the University of London. This organization with a large number of publications has always given rise to various multilevel issues and has inspired other scientists from the area of education research or other areas. They have also established particular software (MLwiN) for multilevel analysis, which is frequently updated to catch up with the new considerations. Their articles cover a wide area of issues in multilevel techniques.

(a) “League Tables”, their limitations and, school performance and comparisons are a major issue of importance in many of their studies. They introduce the use of “Value Added Information” i.e. how many units of improvement an institution has added to its students’ achievement and performance. Such studies are performed by Goldstein & Spiegelhalter (1996), O’Donoghue, Thomas, Goldstein & Knight (1996), Goldstein (1997), Goldstein (1998), Goldstein & Woodhouse (2000), Goldstein, Huiqi, Rath & Hill (2000).

(b) Another issue of interest analyzed in details from the particular group of scientists are applications in the extensions of the basic multilevel models, such as multivariate multilevel models, cross-classification and multilevel repeated measures models. We have already discussed the multivariate case of joint analysis of English and Mathematics Scores (Goldstein et al, 1993). Yang et al. (2002) have elaborated significantly on this issue. Cross-classification in educational data has been analyzed, among others by Hill & Goldstein (1998). Another example of cross-classification of schools in examination results (Goldstein, 1995) is presented here:

The data consist of scores on school leaving examinations obtained by 3435 students who attended 19 secondary schools cross-classified by 148 primary schools in Fife, Scotland (Paterson, 1991). Before their transfer to secondary school at the age of 12 each student obtained a score on a verbal reasoning test, measured about the population mean of 100 and with a population standard deviation of 15. The model is as follows:

$$y_{i(j_1 j_2)} = \beta_0 + \beta_1 x_{1i(j_1 j_2)} + u_{j_1} + u_{j_2} + e_{i(j_1 j_2)} \qquad (4.4)$$

and a number of alternative models were fitted. Random coefficients for verbal reasoning were estimated as zero.

The results of the study are as follows:

- Ignoring the verbal reasoning score, the between-primary school variance is estimated to be more than three times that between secondary schools. The principal reason for this is that the secondary schools are on average far larger than primary schools, so that within a secondary school, primary school differences are averaged. Such an effect will often be observed where one classification has far fewer units than another, for example where a small number of schools is crossed with a large number of small neighbourhoods or a small number of teachers is crossed with a large number of students at level 1 within schools. In such circumstances we need to be careful about our interpretation of the relative sizes of the variances.
- When the verbal reasoning score is added to the fixed part of the model the between secondary school variance becomes very small, the between primary school variance is also considerably reduced and the level 1 variance also.
- When cross-classification is removed by primary school, the between secondary school variance is only a little smaller than in analysis without

verbal reasoning score. Using this kind of analysis, which is typically the case with school effectiveness studies, which control for initial achievement, there were important differences between the progress made in secondary schools. However, most of this is explained by the primary schools attended.

Of course, the verbal reasoning score is only one measure of initial achievement, but these results illustrate that adjusting for achievement at a single previous time may not be adequate.

(c) Although the estimation techniques and algorithms for both the fixed and the random part of a multilevel model were of major importance from the beginning of the adaptation of such models, the rapid development of IT has given rise to this subject. More demanding algorithms and techniques, such as Bootstrap methods, can now be easily applicable in the multilevel software. The "institute of Education" team of scientists has thoroughly dealt with estimation issues. We refer, among others, to the recent works of Browne et al. (2002) and Carpenter, Goldstein & Rasbash (2003).

Multilevel analysis for educational data was introduced by Aitkin (1981). Since then, a great number of articles have been published for this issue, apart from those already mentioned from the 'Institute of Education'. We can refer to Kreft (1993) and McArdle & Hamagami (1994) for simple applications of basic multilevel and logit multilevel models in educational performance. Kreft & Leuew (1998) use basic models multilevel techniques to analyze a large set of data from the National Education Longitudinal Study of 1988 (NELS-88) taken by the National Center for Education Statistics of the US Department of Education. They follow the basic Two-Level approach, however their work has a number of practical advantages. Firstly, they follow a "step-by-step model building" approach which explains perfectly the practical use of Multilevel Analysis. Their critical comments throughout the analysis, stretch some of the major drawbacks in Multilevel Analysis and set the basis for practical considerations. Such problems are estimation techniques and considerations, the effect of "centering" on data in multilevel approach models, 'multicolinearity' issues and so on. Since they compare the results of different subsets of different sample sizes, they illustrate the "sample size selection" issue. Finally, in each step of their analysis, they introduce the basic commands of Mln software (the previous version of MLwiN) as well as references to other known software programs. Afshartous & DeLeeuw (2002) elaborate more on this data set trying to set best techniques for parameter estimation.

In Greece, although the school effectiveness and students' performance issue is of great interest, the only serious attempt to perform multilevel techniques in educational data was taken by Kosmopoulou (1998). In her dissertation project, the author performed multilevel models analysis, and more specifically fitted a 3-level model, in Greek educational data in order to assess school effectiveness and students' performance in the National Entrance Exams of 1990 and 1991 for the student's access in the National Universities and Technical Institutions. The response variable of interest was the mean score of students in the National Entrance Exams. We should notice that according to the educational system being in use at the period of the study, students in the 3rd of Lyceum were divided into 4 scientific orientations ("desmes") and each student according to the scientific orientation should be examined in four subjects in the National Entrance Exams. The explanatory variables were the mean score of students in the third class of Lyceum, the type of school (public/private), the gender of student, the scientific orientation (4 categories) and the year of examination (1990/1991). In the year 1990 the number of level-3 units (prefectures) was 51, the number of level-2 units (schools) was 961 and the number of level-1 units (students) was 52041 while in the year 1991 the corresponding numbers were 51, 978 and 54200, respectively. The author performed a 3-level model and the main final results of the analysis were the following:

- The 3-level model is a significant improvement compared to the simpler 2-level and 1-level models.
- When unadjusted examination results were used (the explanatory variable "mean score of students in the third class of Lyceum" was omitted), girls concluded to perform much better than boys, the type of school was not statistically significant, students of the 1st, 2nd and especially 3rd scientific orientation performed better than those in the 4th and finally, students in 1990 performed better than students in 1991.
- When adjustment was made, it was concluded that boys do better than girls, public schools perform better than the private ones, students who chose the 4th scientific orientation performed better than those in the 1st and finally, again students in 1990 performed better than students in 1991.
- According to the descriptive statistics, the prefecture with the highest mean score in both years was Chios and the prefectures with the lowest mean score were Evros for 1990 and Evritania for 1991.

Although the educational system of National Exams described above is totally different than the system that will be described in our application in the following Chapter, the results of Kosmopoulou (1998) will be the most reliable reference of comparison, due to the lack of other relevant studies in this area of research.

We should also refer to some other projects and results from the Greek literature that will be used for comparison reasons, although their statistical approach is not multilevel modeling but is mainly constrained in descriptive statistics. Some of them are:

- The scaling of the General Admission Grade of the candidate students for the access to the National Universities and Technical Institutions by scientific orientation, published every year by the Greek Ministry of Education, Lifelong Learning and Religious Affairs (www.ypepth.gr).
- The dissertation of Marouga (2004) for the examination of students' performance and preferences according to the Greek National Exams.
- The project of the Centre of Development of Educational Policy of the General Confederation of Greek Workers (GSEE) (2009) about the system of access in the third degree education (National Exams 2004, 2005 and 2006) (www.kanep-gsee.gr/index.php?download=keimeno%206-4-09.doc).
- The report of the Centre of Research and Documentation of the Greek Federation of State School Teachers of Secondary Education (OLME) (2004) expressing the views of the organization for the suppression of the National Exams and the autonomy of Lyceum (http://www.smarinis.gr/aei1.pdf).

4.2 Applications in various areas

In this Chapter we introduce particular examples where various multilevel model techniques are performed in various areas of interests. Such areas have already been described in previous Chapters according to the hierarchical structure of the data measured, and stem from spatial statistics, health research, survey research, repeated measures and meta-analysis.

The scope of this review is to detect in which ways multilevel analysis has introduced in practice in these areas and present the basis of more extended use when multilevel models are proved to be appropriate.

4.2.1 Spatial Statistics

The first article is taken from the area of spatial statistics (Courgeau & Baccaini, 1998).

"Multilevel Analysis in the Social Sciences"

a. INTRODUCTION

The authors used the multilevel approach to study human behaviour taking into account not only individual characteristics but also the fact that these individuals belong to larger geographical units such as communes and regions. This article gives a detailed critical presentation of the models used already and the models to be introduced, as well as the aims and formulations of these models. Attention ranges from the most basic models, which introduce the many different levels in the form of individual and aggregated characteristics, to more complex models, which operate with the random characteristics specific to each level.

Demography has for long favored analysis at aggregated levels. Attention focuses on identifying the relations which exist between the classic demographic rates, corresponding to the phenomenon being studied in each sub-population, and the average values of the characteristics, also calculated for each subpopulation. An analysis of the emigration rates of different regions, for example, would try to link them to the unemployment rates, average incomes, percentage of dependants, etc., found in these regions. The authors state the danger of mistaken inference when we try to infer individual behaviour from aggregated measures ('ecological fallacy) or when we examine individuals separately ('atomical fallacy') and focus on the need to introduce multilevel analysis when referring to emigration models given rise to the idea of working simultaneously on different levels of aggregation, with the aim of explaining a behaviour which is always treated as individual, rather than aggregated as before. Then the authors set up the basic multilevel, as well as the logit multilevel models in a way that has already been discussed in previous chapters.

b. DATA

The example the authors chose to work is that of regional migration in Norway. Data were taken from the Norwegian Population Register to demonstrate the importance of aggregation effects. Various types of model can be used according to the type of data i.e. exponential regression to model the emigration rates of the

regions, logit and event history models to explain the individual risks of migrating in terms of the characteristics of the zones being considered and of the individuals themselves.

Norway has local population registers in which are recorded the demographic events of the individuals living in the country, and in particular their internal migrations (changes of municipal districts). The file used contains the 54814 individuals born in 1958, who lived in Norway in 1991 and who had not migrated abroad. For each of these individuals the successive changes of region (Norway is divided into 19 regions, was known. Only the regional emigration flows, observed over a short period of two years, 1980 and 1981 were considered when the individuals were aged 22-23. A census was conducted in 1980, so various characteristics of the individuals at this date were also known, and there was also the ability to establish how long the individual had been living in the region of residence at the start of 1980. At the individual level eight characteristics had been selected as having a possible effect on the chances of moving out of the region: marital status (married Vs unmarried), being economically active (active Vs non-active), type of occupation (farmer Vs non-farmer), educational level (more than 12 years in full-time education Vs less than 13 years in full-time education), having children (at least one child Vs no children), and the level of income (high income; low income; no income). The authors were then able to reconstitute the aggregated characteristics for the 19 regions (percentage of individuals having left the region in 1980- 1981, percentage of individuals married, percentage of farmers, etc.).

c. ANALYSIS OF INDIVIDUAL AND AGGREGATED CHARACTERISTICS

In the first step, all the analysis was performed separately, either in individual or in regions level (using aggregated data). Although the results were highlighted by the authors we mention that the authors conclude their discussion by: “These early analyses, which need to be further developed, show how understanding of migratory processes can be increased by the simultaneous inclusion in the models of the characteristics of the individuals and of the regions of origin and destination”. They also illustrate the care needed when interpreting the results.

d. ANALYSIS WITH MULTILEVEL RANDOM VARIABLES – DISCRETE RESPONSE DATA

Many demographic characteristics are observed in the form of dichotomous or polytomous variables: an individual is married or not, an individual can migrate between n regions, for example. The authors examined the dichotomous case, the migration flows of the 19 Norwegian regions, for individuals born in 1958 and who migrated in 1980-81. To explain the behaviour they also used the 8 individual and aggregated characteristics that have already defined. Because individual and aggregate characteristics have a specific effect on the probabilities of emigrating from the regions, they introduced them first for each type of characteristic in a simple logit model and in a multilevel logit model. After fitting a number of alternative models for men, the main conclusions were the following:

- The non-random parameters estimated with a multilevel model are in general very close to those we obtain with a simple logit model. But when the effect of the random terms related to the characteristics is not zero at the regional level, a large increase in the dispersion of these parameters is observed, with an approximate doubling of their standard deviation. Despite this, most of the effects that are significant at the 5% threshold in the simple logit model are also significant in the multilevel model. The only exceptions to this rule are two effects of aggregated characteristics: the positive influence that living in a low-income region had on the chances of migrating becomes non-significant in the multilevel model; on the other hand, regions with a high educational level are found to have a significant effect of reducing the chances of migrating when the multilevel model is used, whereas this was not apparent at all with the simple logit model.
- Farmers have a much lower probability of migrating than the other categories, but the higher the percentage of farmers in a particular region, the higher the probability of migrating for all categories. This result highlights the danger of inferring individual results from results obtained at a more aggregated level: the presence of a large number of farmers in a region results in a higher probability of emigrating for all the categories of population, doubtless due to the greater scarcity of non-agricultural employment in such regions. But this does not mean that farmers have a higher probability of emigrating than the other categories, since it is the exact opposite that is observed.

- Individuals with at least one child are found to have a probability of migrating that is always lower than those without children, whether or not we introduce the percentage with at least one child. It is also verified that the variance between regions of the logits of the probability of migrating of those with at least one child is three times higher than that of those with no children when the percentages with at least one child are not introduced. When this percentage is introduced it has a highly significant effect and, above all, it reduces the between-region variances and covariances by half, so that they lose their significant effect: We can therefore say that it does indeed explain a part of these random effects.
- In the case of individuals with more than 12 years of full-time education, they have a higher probability of migrating than the others.
- In a model in which all the characteristics considered are included which have an effect on the regional probability of migrating, the results of a simple logit model compared with a multilevel model show that only the characteristics of educational level are considered to be random between regions.
- The effects of the non-random characteristics are very similar whether the first or second model is used. The case of the farmers now becomes fully significant: the fact of being a farmer always reduces the probability of migrating. By contrast, the fact of having a high income becomes non-significant in a multilevel model.
- The random parameters at the level of the region are modified the most, in relation to the model in which only the fixed educational level characteristic is included. The between-region variance is reduced to a fifth of what it was by the introduction of the other characteristics. On the other hand, the between-region variance remains close to what it was. The correlation between the regional hazards of individuals with more or less than 12 years of full-time education is highly negative whereas it was almost null in the earlier model.
- The final conclusion by the authors is that the introduction of a model using the multilevel random variables does not alter the essential of the conclusions obtained with a simple logit model taking into account the characteristics measured at different levels of aggregation. On the other hand, these random variables provide some valuable information about the relationships between

probabilities of emigrating from the different regions of individuals who have or do not have a given characteristic.

e. CONCLUSIONS

When drawing the final conclusions, the authors state a number of questions to be discussed when a multilevel model is introduced. These are, in brief, the appropriateness and the accuracy of a higher lever measures for human behaviour, the number of levels to be introduced, the need of examination of the behaviours that are specific to each level before we perform a jointed analysis and so on. In any case, however, they recognize the 'rich potential of multilevel models, but also the need to situate them in a coherent theoretical framework'.

4.2.2 Health Statistics

Health statistics is another area where multilevel models can easily be applied due to the clear hierarchy of the data measured in this area (patients are clearly nested within health care providers). Since the similarities with the educational area are obvious, it is not surprise that all the formulas, techniques, algorithms and inferences are borrowed by the educational literature. We have, therefore performance indicators for heath care units analogous to those for educational institutions (Goldstein & Spiegelhalter, 1996) and in general authors from one area to draw conclusions for the other (Goldstein, Browne & Rasbash, 2002). In this chapter we focus on an example taken from the health literature (Carey, 2000).

'A multilevel modelling approach to analysis of patient costs under managed care'

a. INTRODUCTION

The paper is interested in the performance of health care providers and more specifically to analyse the effects of managed care penetration on patient level costs for a sample of 24 medical centres operated by the Veterans Health Administration (VHA) using multilevel modelling techniques. The appropriateness of a two level approach to this problem over ordinary least squares (OLS) is demonstrated. To date, the majority of studies of the determinants of costs have been performed at the hospital level. However, a fundamental practice of managed care associations is

moving patient services out of the hospital inpatient setting to ambulatory sites, medical offices, nursing homes, or patients' own homes. A statistical technique that is well-suited to examining the effects of provider institutions on patients' costs in a managed care environment is multilevel modelling. This framework is used to analyse data that fall naturally into hierarchical structures consisting of multiple 'micro' units nested within 'macro' units. The author clearly emphasises the advantages of a multilevel technique compared to the classical OLS approach, which have been discussed in depth in previous chapters. Also provides the appropriate techniques and formulations for basic multilevel models, as well as for cross-classified models, mentioned but not used in the analysis.

This paper uses a multilevel modelling approach to explain variation in patient costs in facilities operated by the (VHA). The analytic framework developed here explores cost variation occurring at the VHA medical facility level and its interpretation, after controlling for differences in individual patients. The multilevel approach allows for drawing insights regarding the relative performance and efficiency of these public institutions operating in the current environment of managed care.

b. DATA

During 1997, US Veterans Health Administration (VHA) operated 148 medical centres providing a continuum of services under an organized model of managed care in which veterans were gradually being assigned to a primary care physician or team. These institutions provide care for military service connected injuries as well as a broader spectrum of care for poorer veterans. VHA was organized into offering veterans a continuum of treatment settings that in addition to inpatient care include outpatient clinical and surgical services, nursing home care, and home health care. The medical centres were grouped into 22 geographically-based Veterans Integrated Service Networks (VISNs), or groups of medical facilities providing a structural and organizational foundation for integrating services and planning. The structural and functional changes can be summarily characterized as a movement away from hospital based care toward managed care within integrated delivery systems. The data in this study includes 24 medical centres operating in three randomly chosen VISNs and servicing 526117 individual patients in fiscal year 1997.

The primary data sources used in this study come from VHA administrative files. The two level model nests patients within provider units for the fiscal year 1997 (the 1-year period commencing on 1 October 1996).

The response variable is the total direct cost of an individual accrued by VHA during the fiscal year. These costs are adaptations of patient level costs derived from the Cost Distribution Report (CDR), a standard costing procedure in which VHA allocates costs from central accounting sources to specific patient care programmes. The cost of hospital inpatient care, outpatient care, and long-term care are included in the response variable, also including physician costs. The dependent variable used in the analysis is the log of this patient cost.

A critical factor in determining health care cost variation is the clinical state of the patients served. An advantage of multilevel modelling is that it can integrate micro data stemming from a growing volume of patient care information systems. Individual patient health status is measured using the Diagnostic Cost Group (DCG) methodology. DCG models use demographics and diagnoses to predict individuals' relative resource. These may be interpreted as measures of relative health status based on expected expenditure differences in comparison to a benchmark population normalized around a mean of 1.00c. DCGs form risk groups of persons based on diagnoses that are similar. DCG risk scores are well-equipped to control for the clinical status of individual patients in this model of patient annualized costs. Patient level data also includes age and sex, drawn from VHA's Patient Treatment File (PTF) and Outpatient Care File (OPC).

At the medical centre level, the effects of the managed care effort on costs may be related to certain practices that are observable. In order to capture the magnitude of that relationship, a managed care penetration rate variable is included. This is the percentage of ambulatory patients treated at that facility who were assigned to primary care persons or team. As the managed care model is being implemented over time across the agency, the level of primary care oversight achieved at a facility is a means of capturing variation in managed care effort in this cross-sectional analysis. Additional controls at the provider level include size, measured by the number of operating beds (CDR data source), and an indicator for teaching hospitals, gauged by membership in the Council of Teaching Hospitals. Dummy variables for VISN are added in order to control for regional differences in cost such as wage differentials.

Although the author mentions that patients may be classified both by the hospitals they visit and by the physicians they frequent so that individuals within one hospital cluster are not grouped in the same way under physicians (cross-classification), in the VHA this type of cross-classification does not occur, since, in general, physicians and other clinicians operate within a single medical centre. However, patients may receive care from more than one medical centre during the year. This arrangement forms a multiple or cross-unit membership model, a special case of cross-classification.

c. THE RESULTS

Both OLS results and Multilevel Model results were presented for matter of comparisons. The main conclusions drawn from the multilevel models fitted are:

- From the unconditional (null) model, it can be seen that most of the variation in costs occurs at the patient level.
- Adding level-one, the DCG measure is the most powerful predictor of patient costs. Controlling for health status, age is also very highly significant as is sex, with men being more expensive. All of the random effects are also significant, indicating that the relationships between health status and cost and between age and cost differ by medical centre, and adding to the observed strength of the group.
- From the full model results, where the level two predictors are added, these have not 'explained away' the variance at level two: both the intercept and slope variances are still significantly different from zero. Also, the value of the variance component for the age slope coefficient has risen.
- None of the fixed main effects of the level two predictors are significantly different from zero. The managed care penetration rate shows some effect in its interaction with DCG. It is possible that the effect of the managed care variable is operating through its interaction with DCG such that costs rise with DCG risk scores but less so depending upon the extent to which the primary care model has been applied in the facility.
- The significance of the random terms indicates the appropriateness of applying a multilevel model. As anticipated, the standard errors in the multilevel model are much higher than in OLS. A number of coefficients that were significant in OLS are no longer significant in the final mixed model.

- Using the predicted residuals for the intercept and the corresponding C.I., in order to draw comparisons between institutions, the author concluded that there is some difference ib cost containment performance among institutions.
- Judging from the magnitudes of the fixed and random effects, there appears to be greater volatility in cost among institutions across Age than across DCG.
- Using the model coefficients estimates in order to make inferences on patient costs, the author calculates that the quantitative effect of managed care effect on patients cost.

d. CONCLUSIONS

The final conclusions and discussion by the author from the analysis conducted can be summarized into the following:

(a). The author mentions that multilevel statistical modeling provides an important complement to existing econometric techniques in analyzing health care provider costs and efficiency, compared to other techniques such as OLS analysis and panel data models

(b) The author uses the results with caution to perform comparisons between institutions and concludes that 'some deviations were identified and such departures from the average may well serve as useful diagnostics regarding performance evaluation that can inform both local and central management'.

(c) Although the results of multilevel analysis are judged as "inconclusive", the author comments that 'facilities more heavily penetrated by the primary care model appear to be slightly more effective at controlling the costs of their sicker patients'.

(d) The necessity of extensions of the model provided by adding repeated observations on individual providers, as well as more facilities to the data is mentioned, in order to empower the ability of conclusion making by the model.

What can be added in the above conclusions is that the particular article can be used as the basis for research in terms of health care costs for patients, in a variety of health systems, except from the particular one discussed here.

Stemming from the same area of health research and health economics, we refer to an article from Rice & Jones (1997). This article uses all the concepts of the multilevel model already discussed in the thesis (especially from the educational research) 'transformed' into the heath terminology perspectives. The issues described in the article are the basic multilevel model for patients nested within provider units,

parameter estimation for both fixed and random part, residual estimation, some extensions to the basic model, such as cross-classification of patients within provider units and clinicians, and finally, proposals for applications of multilevel techniques in the area of health economics. This review can be seen as an introductory reference for everyone who wants to perform a multilevel analysis in health research.

4.2.3 Repeated Measures

The next example comes from the area of repeated measurements models. Repeated measures are, undoubtedly, a wide area of interest with applications in most of research issues. From the multilevel point of view, the basic characteristic of such models is that individuals form the 2nd level of measurement, where responses/events for the same individual are nested within them, forming the 1st level of analysis. This is exactly the case to be reviewed here (Wright, 1998).

'Modeling Clustered Data in Autobiographical Memory Research: The Multilevel Approach'

a. INTRODUCTION

The article focuses on the issue of autobiographical memory research which involves recording several autobiographical memories for each of several people. These memories are not independent of each other, an assumption of the statistical procedures used in many cognitive psychology papers. To explore autobiographical memory, researchers often ask people to remember several different events. These events will, in some way, be sampled from each person's life. The events are usually different for each person. Conceptually, these events are nested within each person, that is each event happens only for a single person. Therefore the author, compared to all traditional research, states the advantages of the multilevel modeling approach in such research, using examples and concepts from other areas. The four models that are introduced and compared are the simple model that ignores the hierarchy, the aggregated model, the ANCOVA model and finally the multilevel model. Also for the purpose of the analysis the logit multilevel model is introduced.

When studying autobiographical memory, the events and the memories are unique for the person and this is important for theories of autobiographical memory.

The events are nested within the person in a manner analogous to the basic hierarchical structure in multilevel modeling.

b. DATA

The example used here is an example of autobio- graphical memory research by Burt et al. (1995). At the beginning of the summer holiday, 27 volunteers were given several rolls of film with the instructions that they should take photographs as they normally would. Several thousand photographs were taken. Some were excluded because they were out of focus or one of several depicting the same event (for example, many pictures of a wedding ceremony). Others excluded for missing values. The number of photographs kept for analyses is 1342. The number per person ranged from 23 to 147. Photographs are nested within people. The photographs were coded by whether they were of an activity, of participants, of a location, or some combination of these.

After the summer holiday, these photographs were presented to subjects for 10 seconds on a tachistoscope. Several distractors, taken from the researchers' own collections, were also presented. Over 97% of them were correctly recognized as foils. Subjects were asked if they could remember the event and the reaction times were recorded. If they could remember the event, they made several ratings about it, including the importance of the event and the level of emotion provoked by the event at the time. If they could not remember the event, they were asked whether they thought it was a distractor, whether there were too many similar events to be sure, or whether the picture did not provide sufficient cues for them to decide if it was their Only three relationships were examined by the author since they cover three of the main statistical procedures used by cognitive psychologists. The first is comparing reaction times by whether the event was correctly recognized, or not, and if not recognized the reason they gave. The second is the relationship between importance and emotion. Both variables were measured on 7-point rating scales and the Pearson correlation was found in the original analyses. The final relationship is between the type of photograph (of activities, participants and/or locations) and whether the subject could remember the event.

c. REACTION TIMES

There was a missing value for one reaction time, making the n=1341 for these analyses. The reaction times were highly skewed. There were some that were longer than 10 seconds, which was the duration that the picture was presented on the tachistoscope. These cases were recoded as 10 seconds. The natural logarithms of these data were taken so these transformations removed most of the skewness and made the distribution appear roughly Normal. The response variable was called 'lntime'. About 65% of the holiday photographs were remembered (n=879). This will serve as the baseline category, with dummy variables for distractors (`dist' n=40), those not recalled because there were others that were too similar (`simil' n =222) and those for which the cues were insufficient (`cues' n =200). Besides `remembered' being the most frequent category, and therefore providing more reliable parameter estimates, the most logical contrasts are between `remembered' and each of the others. The final approach for this model is multilevel modeling which allows the intercept, here the times for remembered photographs, to vary among subjects but in addition it assumes that these response times are Normally distributed.

The results showed that the reaction times are faster for the remembered events and moreover that the variance of the subject level residuals is significant. In other words, that it is important to take into account the variation among subjects and the assumption of independence is not valid. Also, by fitting random variables in the random part of the model, it was shown that the variance for the non-remembered photographs was higher than for the remembered photographs.

d. IMPORTANCE AND EMOTION

In this step the relationship between the importance of an event (import) and people's emotional reactions to the event (emot) is examined

Subjects only made ratings for the events they remembered. Of the 879 remembered photographs, there was one missing value for emotion reducing the number of photographs to 878. Both ratings were on 7-point scales and they were treated as continuous interval measures

The multilevel approach treats photographs as a random sample from each person's summer but also treats subjects as a random sample from some population. The author fits the model with both random intercept and random slope and the results show that emotion gets higher when importance gets higher, as well as, the variation

among subjects is largest when importance scores are either high or low, since the relationship between importance and subject variance is quadratic.

e. PREDICTING REMEMBERING

In the last part of the analysis the author tried to determine the best cues for retrieving memories. In other words, to predict whether a person remembers the event, using three dummy variables for whether a participant, a location and/or an activity are depicted in a photograph.

In the multilevel approach a logistic multilevel model was used as follows:

$$\log it\pi_{ij} = \beta_0 + \beta_1 part_{ij} + \beta_2 act_{ij} + \beta_3 loc_{ij} + u_j + e_{ij} \qquad (4.5)$$

where $part_{ij}$, act_{ij} and loc_{ij} are the three dummy variables as defined before and π_{ij} is the probability of remembering. The model provided a relatively good fit and moreover the assumption of binomial variation in the residuals of photographs was not rejected.

f. CONCLUSIONS

The author concludes once again by mentioning the advantages of multilevel techniques when analyzing autobiographical memory data. Also the advantages in other areas of psychology, where events are nested into individuals. It is pointed out, however that, in contrast to other areas, where the hierarchy is profound, 'it is perhaps less intuitive that memories are nested within a person'.

4.2.4 Survey Research

The next two examples to be reviewed come from the area of survey research which is a wide area for applications for multilevel modeling techniques. The first (Rice et al, 1998) examine a clear structure of hierarchy where individuals are nested within households, which are also nested within geographical areas of residence. In the second case the effects of interviewers and respondents on the data collected from a survey research are examined (Hox, 1994).

1. 'The Influence of Households on Drinking Behaviour: A Multilevel Analysis'

a. INTRODUCTION

The purpose of the analysis is to examine the influence of household membership and area of residence on individual drinking behaviour. Before any statistical analysis a thorough theoretical review points out, one the one hand, the dangers of long heavy drinking compared to the intermediate drinking, and on the other examines the factors that affect drinking consumption and drinking behaviour of an individual. Using a large number of examples and references, the authors point out that grouping of individuals affects their drinking behaviour. Such groups can be of many kinds (social, religious, geographical) and can affect either in a positive or negative way. Households in which individuals are naturally nested are said to be one of the fundamental factors that affect the individual's behaviour as well as the area of residence. It is also discussed that such groups tend to make individuals that belong to them more "homogeneous" that individuals nested in different groups. Due to this hierarchy, the authors introduce and suggest multilevel model for data analysis and also introduce the basic models and parameter estimations, already explained in previous chapters.

b. DATA

Data from the 1993 Health Survey for England (HSE) was used for the study. Collection was performed throughout 1993 and on into early 1994, and consists of 17687 interviews with adults (aged 16 or over) living in 9700 households in England. The sample was distributed relatively evenly across the 14 English Regional Health Authorities and was obtained by sampling households from 2 or 3 electoral wards of residence in each Authorities' area. The survey is unusual in seeking responses from all adults in each household, providing a rare opportunity to explore effects within households. The 1993 survey focused on cardiovascular disease and associated risk factors, including alcohol consumption, as well as general health and various long-standing illnesses. There are well-known problems with the measurement of lifestyles in household surveys, associated with under-reporting and low response rates of heavy drinkers (Warner, 1978). Nevertheless, such information sources remain the method by which governments monitor their success in reaching drinking targets and are the only datasets large enough to answer the sort of questions addressed in this

study. Alcohol consumption data was incomplete for 1139 individuals and these were excluded from the analysis. A further 1119 individuals were also excluded due to missing responses to the various explanatory variables used. This resulted in a total of 15429 individuals within 8737 households within 495 enumeration districts presenting for analysis. One-person households were retained in the analysis as they contribute to the estimates of the covariates of alcohol consumption and to area variations in consumption. The vast majority of the sample live in two-person households. The following variables were included in this analysis:

Personal characteristics: Gender, age.

Social environment/support: No. of persons in household, whether single or have a partner, perceived social support.

Health: Perceived stress.

Health related activity: Physical activity level.

Educational: Educational attainment.

Socio-economic: Social class, car ownership, whether or not economically active, whether in receipt of income support.

Various other potential explanatory variables were available in the HSE, for example, smoking status and self-reported general health. However, they were excluded since the relationship between these variables and drinking status was likely to be simultaneously determined.

The dependent variable in this analysis was an estimate of the number of units of alcohol drunk in a 1-week period based on respondents' answers to questions relating to the frequency of consumption and the number of units consumed on a usual occasion (Bennett, 1993). The response data displayed strong skewness and transformation to a more symmetrical distribution was sought by taking the natural logarithm. The distribution of alcohol consumption also often contains a significant proportion of zero values. Because of the problem of zero observations, a constant of unity was added to all observations before taking natural logarithms.

Empirical analyses of individual behaviour incorporating a household effect relied on the use of an explanatory variable often in the form of a dummy variable indicating whether or not other household members drink (or drink heavily), or a continuous measure of the average units consumed per individual within the household

The approach adopted here was to model both area and household effects as random components within a multilevel framework. To investigate the effects of area of residence and household membership on individual drinking behaviour a multilevel model including random components for individual, household and geographical area was specified. All explanatory variables were entered as dummy (0,1) variables except age, which is continuous. To ensure variations at each of the three levels were estimated at typical values of age, age was centred about its mean of 46 before being placed in the models.

c. RESULTS

After fitting the appropriate models, the results can be summarized as follows:

- Drinking declines with age while levels of consumption amongst the younger age groups are moderately high. In the case of women drinkers aged between 16 and 34 the average alcohol consumption was moderate as well as for men of the same age, but with higher average values for men.
- Males generally consume more alcohol than females and there is a quadratic age effect indicating that older people drink less. There is also an indication that there is a differential age effect for males and females shown by the significance of male by age interaction terms. It appears that although males drink more than females, the difference decreases with age. Single people tended to consume more alcohol.
- Of the general health and activities characteristics, individuals who report moderate and vigorous activities generally consume more alcohol compared to their baseline category of inactive individuals. There is no evidence in these data to suggest that a lack of social support or increased stress has an effect on levels of alcohol consumption. Individuals who are unemployed, inactive or working part-time are less likely to drink heavily than those engage in full-time employment (base-line category). However, the effect observed for inactive respondents decreases dramatically when household size and car ownership are included. This suggests that in the ``individual model'' the economically inactive was inappropriately picking up, an effect that should be properly attributed to household-level variables
- As for household characteristics, there was a clear household size gradient indicating strongly that larger households are associated with decreased

individual alcohol consumption. Car ownership is likely to be a reflection of individual current and capital household wealth and ownership of multiple cars appears to be associated with increased alcohol consumption.

- In all the models presented, the estimated variation at each of the three levels appears as statistically significant.
- The majority of the variance is attributed to differences between individuals (56%). However, 42% of variation occurs at the household level indicating that household membership and composition is very influential in determining inhabitants' consumption levels. Very little of the variation is attributed to area effects (2%), and speculation of strong geographical contextual effects of drinking behaviour appears unfounded in these data.
- There is a large amount of unexplained variation in individual alcohol consumption, which can be attributed to household membership. Further, little variation is attributed to differences in geographical area influences. The influence of household membership is nearly as great as that due to differences between individual characteristics in determining consumption of alcohol.

d. CONCLUSION

Conclusions of the author are mainly focused on suggestions about policies to reduce drinking consumption taking into account, of course, the results of the analysis. It is pointed therefore that focus will be given on both individuals and households, since their 'contribution' in drinking behaviour is, more or less, equivalently important. Geographical contextual effects were found to be minimal, however the authors mention that the definition of 'area' used in the analysis might be inappropriate.

2. 'Hierarchical Regression Models for Interviewer and Respondent Effect'

a. INTRODUCTION

The example, reported by Hox (1994) concerns the effect of interviewers and respondents on survey results. Both respondents and interviewers have been recognized as a potential source of error (observational error) in survey interview data. Interviewer and respondent characteristics can have an important effect on the survey results and quality, and much methological research has been spent on the

question how much interviewer and respondent bias is present in social survey data. Since respondents are nested within interviewers, in methological terms, this is clearly a multilevel problem. The specific example investigates how much interviewer and respondent characteristics influence the speed of interviewing (i.e. how many questions have been asked and answered in a given time period) All the traditional methods are mentioned, however the analysis is performed by applying hierarchical regression model techniques.

b. DATA, DATA ANALYSIS & RESULTS

The example data stem from a controlled field experiment on mode effects (De Leeuw 1992). In this example, data are analyzed from 515 respondents, who were questioned by 20 interviewers. Three data collection methods are compared: 221 of the interviews were conducted face-to-face, 219 by telephone using a pencil-and-paper questionnaire, and 75 by telephone using Computer Assisted Telephone Interviewing (CATI), all three using the same interviewers. The respondents were randomly assigned to the different collection methods: in both telephone conditions, they were randomly assigned to interviewers. Due to financial constraints, in the face-to-face condition, random assignment of respondents to interviewers was used within four broad geographical regions.

The dependent variable in the analysis is the total time needed for an interview. Because time measures generally have a skewed distribution, an inverse transformation is used, which transforms the variable, time, into the variable speed. Thus the dependent variable Y_{ij} is the speed of the interview measured in number of questions completed per minute. The research problem is whether interviewers differ in the speed with which they complete an interview. In addition, the author wishes to analyze which interviewer and/or respondent characteristics influence the interviewing speed. The explanatory variables $X_{pij}^{(1)}$ at the respondent level include two dummy variables indicating the three data collection methods: one contrast variable, tel (coded +1, -1), that compares the two telephone conditions to the face-to-face condition, and one contrast variable, cati, that compares the CATI-condition with the pencil-and-paper telephone condition (cati). The other respondent variables are respondent age (r-age) and loneliness (lonely), a measured by a multi-item scale. The explanatory variables $X_{qj}^{(2)}$ at the interviewer level are amount of previous

interviewing experience, interviewer age (i-age), interviewer preference for telephone interviewing (pref.tel) and the interviewer's score on five personality scales: extroversion (extro), friendly disposition (friendly), conscientiousness (cons.), social assurance (soc.ass.), and ability to terminate awkward situations (term.).

Because the design is not completely orthogonal, the first step in the analysis is to inspect the correlations between respondent and interviewer explanatory variables. The correlations between respondents and interviewers are generally low, indicating that the partial othogonalization was successful. But, because the respondent and interviewer effects to be investigated are generally also small, it is safer to take these correlations into account in the analysis by modeling the interviewer effects conditional on the respondent variables.

The starting point for the model construction is the intercept-only model, which is a model with no explanatory variables. It is given by the following equation:

$$y_{ij} = \beta_0 + (u_{0j} + e_{ij}) \qquad (4.6)$$

This model contains the fixed regression coefficient β_0 for the grand mean for y_{ij} and the two variance estimates, σ_e^2 for the residual variance at the respondent level and σ_{uo}^2 for the residual variance at the interviewer level. In the example, β_0 is estimated as 3.19, indicating an overall interviewing speed of slightly more than three questions per minute. The respondent level variance σ_e^2 is estimated as 0.68 and the interviewer level variance σ_{uo}^2 as 0.11; this model produces an estimate for the intra-interviewer correlation ρ_1 of 0.14.

The next analysis step analyzes explanatory variables at the lowest (respondent) level as fixed variables; that is, without the corresponding variance components for the regression slopes. Equation is now:

$$y_{ij} = \beta_0 + \beta_p X_{pij}^{(1)} + (u_{0j} + e_{ij}) \quad (4.7)$$

In the subsequent model, the regression coefficients of the respondent variables are assumed to be random; that is they are assumed to vary between interviewers. This is described by the following equation:

$$y_{ij} = \beta_0 + \beta_p X_{pij} + (u_{pj} Z_{pij} + u_{0j} + e_{ij}) \qquad (4.8)$$

In equation (4.8), each regression slope β_p, has a corresponding random error term $u_{pj}Z_{pij}$. In the example data, the effect of the CATI contrast turns out to be not significant (p=0.25). The other explanatory respondent variables are all significant. Only the regression slope for the telephone contrast has a significant variance component. The conclusion is that the model for the correspondent effects might be simplified by dropping the CATI contrast altogether and assuming a random slope only for the telephone contrast. In this model, the respondent level variance is 0.53, and the interviewer level intercept variance is 0.08.

The next analysis step adds the explanatory variables at the interviewer level, giving

$$y_{ij} = \beta_0 + \beta_p X_{pij}^{(1)} + \beta_q X_{qj}^{(2)} + (u_{pj}Z_{pij} + u_{0j} + e_{ij}) \quad (4.9)$$

In the example, only three of the nine interviewer variables are significant; interviewer training, preference for telephone and extroversion.

The between-interviewers variation of the regression slopes for the telephone condition can be modeled by including interactions between the telephone condition variable and explanatory variables at the interviewer level. This gives the full model, which can be formulated as

$$y_{ij} = \beta_0 + \beta_p X_{pij}^{(1)} + \beta_q X_{qj}^{(2)} + \beta_{pq} X_{qj}^{(2)} X_{pij}^{(1)} + (u_{pj}Z_{pij} + u_{0j} + e_{ij}) \quad (4.10)$$

The only significant interaction effect is the interaction of the telephone contrast with the interviewer variable, social assurance. Because the interpretation of interaction effects requires that the corresponding simple effects are also included in the model, the (nonsignificant) interviewer variable, social assurance, is again added to the model. To aid interpretation, social assurance is centered around its overall mean of 61.8, and the interaction term is computed using the centered variable. In model (4.6), the residual variance at the respondent level is 0.68, and the residual variance at the interviewer level is 0.11. In model (4.8) σ_e^2 is estimated as 0.53; this can be expressed by saying that the respondent variables reduce the residual variance at the respondent level by 23%. Similarly, the interviewer variables can be said to reduce the residual variance at the interviewer level by 22%. The explanatory interviewer variables added in model (4.9) reduce the intercept variance by a further 43%. Adding the (nonsignificant) interviewer variable, social assurance, and its interaction with the telephone contrast (model (4.10)) reduces the intercept variance by 3% and the

variance of the regression slope for the telephone contrast by 22%. Interpreting these variance reductions as explained variance, a comparison of the explained variance across the different models fitted shows that both the respondent and the interviewer variables explain a significant portion of the initial variance in interview speed. The interaction that is added in model (4.10) does not appear to explain much variance but, in fact, does explain a considerable proportion of the slope variance that appears in the previous model (4.9).

Calculations, however, of the explained variance, by using simply the variance components and residuals in a model with random slopes should be performed with caution. This is mainly because the variance components are not generally invariant under admissible linear transformations of the explanatory variables.

Using the models' deviances for a chi-square test shows that, in all comparisons of consecutive models, the more complicated models have a significantly better fit. Most of the regression coefficients are stable between different models. Although interviewer and respondent variables are correlated, adding the interviewer variables to the model does not appreciably change the regression slopes for the respondent variables. Only the intercept changes. The interpretation of the regression slopes is straightforward. Interviews take longer with older and lonely respondents, previously trained and extrovert interviewers are faster, and interviewers that have expressed a preference for using the telephone are also faster. The regression contrast for the telephone condition is coded –1 for the face-to-face condition and +1 for both telephone conditions. Its slope coefficient of 0.30 means that the telephone condition is faster by (2 x 0.30=) 0.6 questions per minute; at an overall average of 3.19 questions per minute, this means that telephone interviews are 19% faster. However, because the variable, telephone condition, is involved in an interaction, the interaction effect and the corresponding simple effects cannot be interpreted in isolation. When an interaction between two explanatory variables is involved, the simple regression coefficients for either of these variables reflect a conditional relationship, which is the relationship that holds when the other explanatory variable has the value zero. Because social assurance is centered around its overall mean of 61.8, the regression slope for the telephone contrast reflects the effect of this explanatory variable for interviewers with an average social assurance. To interpret the interaction, the author concludes that over the observed range of values for social assurance, telephone interviewing is faster than face-to-face interviewing. In the telephone interview, there

is no relationship between social assurance and interviewing time, but in the face-to-face interview, interviewers with a higher social assurance tend to use more time. For an explanation, it could be hypothesized that the more personal situation in the face-to-face interview leads the less socially assured interviewers to adopt a task-orientated role, whereas the more socially assured interviewers adopt a social role, which uses up more time. In the more businesslike situation of the telephone interview, this differential role assumption does not take place.

c. CONCLUSIONS - REMARKS

The author ends with some critical conclusions, which can have both theoretical and practical affect on the area of survey research.

- Instrument effects, such as the type of questionnaire, is suggested to be a new level of analysis in a multilevel model. To do so, different type of questions could be treated as repeated measures within respondents i.e. all types could be asked to all respondents.
- Even though the interviewer effect is not of practical important for a survey (not included in the results), it should be measured in the hierarchical analysis as a control factor. Even a small intra-class correlation can cause large bias in the parameter estimates and therefore lead to misleading results
- Ideally, respondents should be assigned randomly within interviewers, otherwise respondent and interviewer characteristics could be confounded. However, in survey research companies, where this is not a frequent situation, multilevel techniques could solve the problem of confounding since respondent variables would be controlled for interviewer effects and vice versa.

Although survey research appears to be an area where multilevel techniques could be readily used, both in theory and in practice, compared to other research areas such as educational research, these techniques are still rarely preferred to the classical techniques, especially in survey research companies.

4.3 Conclusions of the Chapter

The main conclusion which can be drawn from the discussion of this Chapter is that Multilevel Analysis Techniques are extensively applied from a great range of scientists and researchers, not only in the area of Educational Statistics, where the hierarchical structure of data is profound, but also in less profound research areas such as spatial statistics, health research, survey research, repeated measures and meta-analysis. Moreover, as discussed also by the authors, the use of multilevel instead of more classical statistical techniques has significant advantages both in the precision and the statistical accuracy of the results, as well as in the interpretation of the hierarchical structure of the data. However, in order to enhance more on the advantages of the multilevel techniques, in the following Chapter we will perform a practical application of Multilevel Analysis in a real Greek educational dataset referring to the General Admission Grade of students to the National Exams.