Journal of Official Statistics, Vol. 20, No. 3, 2004, pp. 481–493

On the Distribution of Random Effects in a Population-based Multi-stage Cluster Sample Survey

Obioha C. Ukoumunne, Martin C. Gulliford, and Susan Chinn[1]

Data from surveys are often characterised by clustering of individual level responses within higher level units such as households, enumeration districts or counties. Multilevel modelling is an appropriate method for analysing data from such studies, but an assumption of normality is required if the estimated standard errors are used to make inferences about the parameters. We evaluated the distribution of random effects at the postcode sector and district health authority levels of clustering for 13 health outcomes and lifestyle risk factors using data from the Health Survey for England 1994. Normal plots supported the assumption of normality for eight outcomes. A positive relationship was found between skewness at the individual and cluster levels. The findings of this study suggest that for outcomes with non-normal distributions at the cluster level, the application of a normalising transformation to the individual level residuals may also have a normalising effect at the cluster level.

Key words: Multilevel models; hierarchical data; components of variance.

1. Introduction

Data from sample surveys and intervention studies are often characterised by an hierarchical structure (Rice and Leyland 1996; Duncan, Jones, and Moon 1998). Hierarchical or clustered data structures are generated by studies that utilise multi-stage cluster sampling, which is typically employed in large complex surveys. Clustered data may also be encountered in intervention studies when cluster randomisation, rather than individual randomisation, is used. Cluster sampling may be implemented for reasons of cost-effectiveness and to avoid the need to obtain a sampling frame of individuals (Moser and Kalton 1971). Cluster randomisation may be used for a variety of reasons such as to avoid contamination between intervention groups and to evaluate interventions that are naturally applied at the level of area or organisation (Donner and Klar 2000). Even when an hierarchical structure is not imposed by the design, subjects are typically clustered or nested within organisation units such as hospitals, workplaces and schools and administrative areas such as health authorities, counties and electoral wards, as in the United Kingdom.

When analysing individual level outcomes from datasets with an hierarchical structure, allowance needs to be made for the correlation between the responses of subjects who share the same cluster. Subjects living in the same community or treated in the same

[1] Department of Public Health Sciences, King's College London, 5th Floor, Capital House, 42 Weston Street, London SE1 3QD, U.K. Email: obioha.ukoumunne@kcl.ac.uk

Acknowledgments: Data from the Health Survey for England 1994 are Crown Copyright. They were made available by the Office for National Statistics through the Data Archive and have been used by permission. Neither the Office for National Statistics nor the Data Archive bear any responsibility for the analysis or interpretation of the data reported here. Obi Ukoumunne is supported by an MRC Special Training Fellowship in Health Services and Health of the Public Research. We are grateful for the valuable comments of three referees.

hospital, for example, will be subject to environmental and contextual influences that may effect on their health. Clusters may also tend to attract subjects with particular characteristics. If those characteristics are related to the outcomes of interest then one might expect the outcomes of patients sharing the same cluster to be correlated. This correlation within clusters gives rise to an additional source of outcome variation, that is, between clusters. The assumption of independence underlying the use of standard statistical methods is not met under these conditions and alternative approaches must be used. The multilevel model, which allows for the within cluster correlation between subjects' responses, is appropriate for the analysis of such data (Goldstein 1995). The multilevel model allows parameter estimates in regression models to vary randomly between cluster units, and in so doing explicitly models the correlation in observations taken from the same cluster. Under the simplest multilevel model the responses of subjects in the same cluster are treated as being partly determined by a random effect unique to that cluster. The random effects, or residuals, for a given level of clustering are assumed to come from a common distribution with mean zero, with each representing the extent to which the mean response for the cluster differs from the overall mean response. The variance of these effects represents the natural variation between clusters on the outcome of interest.

Although normality is not required to obtain consistent estimates under the classical multilevel model, the use of standard errors to make inferences does rely on the assumption that the random effects are drawn from a Normal distribution (Goldstein 1995). The validity of this assumption may be hard to test at the cluster level because the number of clusters in surveys and trials is often too small to quantify accurately the degree of departure from normality (Feng et al. 2001). Yet it is for studies with small numbers of clusters that departures from the assumption of normality are of greatest importance. Turner, Omar, and Thompson (2001), using a Bayesian framework, showed how incorrectly assuming normality in the distribution of cluster level effects may lead to influential cluster units being inappropriately weighted and regression coefficients being incorrectly estimated.

Empirical information on the distribution of cluster level random effects may provide evidence on the extent to which outcomes can be expected to be normally distributed. This study, using data from the Health Survey for England 1994, describes the distribution of cluster level random effects for several continuous health outcomes and lifestyle risk factors. It also explores the relationship between the distributions of residuals at the individual and cluster levels and investigates the extent to which non-normality at the cluster level may be reduced by using transformations that normalise the individual level residuals. In the next section the Health Survey for England 1994 is described. Section 3 describes the methods used to estimate the cluster level random effects and summarises the distribution of these for several continuous outcomes from the survey. Section 4 discusses the implications of the findings.

2. The Health Survey for England 1994

Data from the Health Survey for England 1994 (Crown Copyright 1994) were obtained from the Data Archive, University of Essex, Essex, England. The Health Survey for

England is an annual cross-sectional survey of health and lifestyle with a multi-stage random sampling design, covering adults living in private households in England. It is carried out by the Joint Health Surveys Unit on behalf of the United Kingdom Office for National Statistics. A full description of the design of the survey is available (Colhoun and Prescott-Clarke 1996). The survey used postcode sectors as the primary sampling units and households as the secondary sampling units. The sampling of postcode sectors was stratified by health region, the proportion of people aged 65 and over, the proportion of households without a car, the proportion of economically active males who were unemployed and the proportion of non-white adults. A random sample of postcode sectors was drawn using the Postcode Address File as the sampling frame with the probability of selection proportional to the number of addresses. Eighteen addresses were randomly sampled from each postcode sector. Some addresses contained more than one household. When this occurred a maximum of three households was sampled at a given address. All subjects 16 years and over within the selected households were eligible for the study. Data were collected from the survey respondents through structured face-to-face interviews. Physical measurements and blood samples were taken by a nurse on a separate visit. The survey was conducted such that interviews and nurse visits in each quarter of the year were conducted with fully representative subsets of the total sample. Smaller numbers of people were willing to see the nurse than to receive the structured interview, so that 71% of all individuals estimated to be eligible were interviewed but only 62% responded to the nurse visit and blood samples were obtained for only 51%.

Postcode sectors are groupings of postcodes used to classify individuals according to geographic location of residence and thus to aid the delivery of mail in the United Kingdom. Postcodes are also the primary geographic unit used for recording vital statistics and health service activity in Britain (Donaldson and Donaldson 1993). One of the factors used to stratify the sampling of postcode sectors, health region, was also a clustering variable in the dataset. Each health region was managed by a regional health authority (RHA) responsible for strategy, allocation of resources and monitoring performance (Donaldson and Donaldson 1993). Information was also recorded on the district health authority (DHA), another type of cluster, within which the postcode sectors were situated. District health authorities were the principal purchasers of health care in the United Kingdom National Health Service at the time of the survey and played a more direct role in the delivery of health care than the regional health authorities. Their roles included the assessment of health needs of the local population, prioritising those needs and organising contracts to secure services for meeting them (Donaldson and Donaldson 1993). Although DHAs were not explicitly recognised in the sampling design, district health authority level analyses were an important output of the survey and these were incorporated into national health indicator datasets. The Health Survey for England 1994 dataset thus had a five-level hierarchical structure with individual subjects nested within households nested within postcode sectors nested within district health authorities nested within regional health authorities. In 1994 the survey covered all 14 regional health authorities and 177 district health authorities, 720 of the 7,223 postcode sectors, 9,068 households (77% of those eligible) and 15,809 individuals (71% of those estimated to be eligible). Typical regional health authorities, district health authorities and postcode sectors had total populations of 3 million,

250,000 and 6,500 respectively, at the time of the 1994 survey. Approximately 80% of the total population in England is aged 16 and over.

3. Distribution of Cluster Level Random Effects

3.1. Methods

MLwiN Version 2.1 (Rasbash et al. 2000) was used to fit multilevel models to and calculate cluster level random effects from the Health Survey for England data. An identical model was fitted to the dataset as in a study of between cluster variation in health outcomes (Gulliford, Ukoumunne, and Chinn 1999). A five-level variance components model was used to analyse the outcomes, allowing for variation between clusters at the regional health authority, district health authority, postcode sector and household levels. The fitted model was:

$$y_{ijklr} = \mu + \alpha_i + \beta_{j(i)} + \gamma_{k(ij)} + \delta_{l(ijk)} + \varepsilon_{r(ijkl)}$$

where y_{ijklr} is the response for the rth subject in the lth household, the kth postcode sector, the jth district health authority and the ith regional health authority; μ is the overall mean; α_i, $\beta_{j(i)}$, $\gamma_{k(ij)}$, $\delta_{l(ijk)}$ are the random effects associated with the ith regional health authority, jth district health authority, kth postcode sector and lth household respectively; and $\varepsilon_{r(ijkl)}$ is the rth residual effect within households. Brackets are used in the subscripts of random effects terms to indicate the nested structure of the data. For example $\beta_{j(i)}$ is the random effect of the jth district health authority nested within the ith regional health authority. The α_i represent the extent to which the mean of the ith regional health authority differs from the overall mean; the $\beta_{j(i)}$ represent the extent to which the mean of the jth district health authority differs from the mean of the ith regional health authority within which it is nested, and so on. Effects at each level are independent, conditional upon the random effects at higher levels. They have zero mean and associated components of variance σ^2_{RHA}, σ^2_{DHA}, σ^2_{PCS}, σ^2_{HH} and σ^2_e, respectively.

The Restricted Iterative Generalised Least Squares (RIGLS) estimation procedure was used to fit the models. Standardised residuals at the levels of individual subject, postcode sector and district health authority were saved. MLwiN applies a shrinkage factor to the raw residuals at the clustering levels such that they are decreased in absolute value. Shrinkage is used to reflect the relative lack of information provided by cluster units that contain smaller numbers of subjects. The shrunken residuals theoretically provide a more accurate representation of the underlying distribution of random effects than the raw residuals.

Thirteen continuous outcomes were analysed: serum cholesterol (mmol/litre); glycated haemoglobin (%); plasma fibrinogen (g/litre); serum ferritin (μg/litre); haemoglobin (g/dl); systolic blood pressure (mmHg); diastolic blood pressure (mmHg); body mass index (kg/m^2); waist circumference (cm); hip circumference (cm); General Health Questionnaire score (Goldberg 1972); number of units of alcohol drunk per week; and the mean number of cigarettes smoked per day. Some of the outcomes were analysed both as untransformed variables and after applying normalising transformations. The "ladder" command in the Stata 7 software (StataCorp 2000) was used to identify suitable

transformations. The "ladder" command tests a selection of power transformations (Y^3, Y^2, Y, $Y^{1/2}$, $\log(Y)$, $Y^{-1/2}$, Y^{-1}, Y^{-2} and Y^{-3}) to assess whether they convert a variable, Y, into a normally distributed one. The command applies a Chi-squared test based on Royston's (1991) adjusted version of D'Agostino, Belanger, and D'Agostino's (1990) test for non-normality. For many transformations the Chi-squared statistic was so large that it was not presented in the Stata output. The approach used in this study was to choose from among the transformations for which the p-value of the Chi-squared statistic was given, as these were the only plausible transformations. Histograms and normal plots were then used to select the best transformation from amongst those considered plausible.

It is acknowledged that transformations that normalise the raw outcomes themselves do not necessarily also normalise the individual level residuals since the outcome is partly composed of cluster level effects. In this study, however, the outcome and the distribution of individual level random effects did have nearly identical distributions, so it was possible to assess the extent to which transformations that normalised the individual level residuals also normalised the distribution of cluster level random effects.

3.2. Findings

Normalising transformations were identified for nine outcomes: a reciprocal transformation was chosen for glycated haemoglobin, systolic blood pressure, body mass index and hip circumference; a log transformation was chosen for serum ferritin and diastolic blood pressure; a squared root transformation was chosen for plasma fibrinogen; a squared transformation was chosen for haemoglobin; and a reciprocal of squared root transformation was chosen for waist circumference. The reciprocal and reciprocal squared root transformations were equally normalising for body mass index and the former transformation was chosen on the basis that regression coefficients may be back-transformed to a more commonly used quantity. In practice, when analysing data from a study, the same transformation would be used to analyse systolic and diastolic blood pressure (Solomon 1985), but different transformations are used here since each was simple and effective with respect to normalising the respective outcomes. None of the power transformations for serum cholesterol, General Health Questionnaire score, number of units of alcohol drunk per week and mean number of cigarettes smoked per day were close to normality. The normal plots indicated that there were no powers larger than 3 or smaller than -3 that would provide normalising transformations for these variables. The large proportion of subjects scoring zero on the GHQ score, units of alcohol drunk per week and cigarettes smoked per day meant that no continuous transformation could make them normal. The normal plots for the transformations of serum cholesterol showed that the identity transformation was positively skewed and that although the squared root transformation reduced the skewness, it increased the excess kurtosis. Fractional powers between 0.5 and 1 were tested, but none produced distributions that were noticeably more normal than the untransformed outcome.

In total 13 variables and 9 transformed variables were analysed. Table 1 summarises the components of variance for each variable. For fibrinogen and its squared root transformation the variance component at district health authority level was zero and

Table 1. Components of variance at the individual, household, postcode sector (PCS), district health authority (DHA) and regional health authority (RHA) levels

Outcome	Individual level	Household level	PCS level	DHA level	RHA level
Serum cholesterol, mmol/litre (untransformed)	1.33073	0.26368	0.03944	0.00562	0.00289
Glycated haemoglobin, % (untransformed)	0.94336	0.16659	0.03919	0.00031	0.00023
Glycated haemoglobin, % (reciprocal)	0.00033	0.00008	0.00002	8×10^{-9}	1×10^{-20}
Plasma fibrinogen, g/litre (untransformed)	0.47097	0.17961	0.03647	0	0.00035
Plasma fibrinogen, g/litre (square root)	0.03603	0.01330	0.00278	0	0.00004
Serum ferritin, μg/litre (untransformed)	6,638.44	0	35.2271	2.98164	1×10^{-20}
Serum ferritin, μg/litre (log)	0.74044	0	0.00494	0.00014	0.00071
Haemoglobin, g/dl (untransformed)	1.80010	0	0.03367	0.00708	0.00099
Haemoglobin, g/dl (square)	1,374.34	0	25.2060	4.84925	0.96282
Systolic blood pressure, mmHg (untransformed)	272.107	123.837	10.8804	3.67632	1.50369
Systolic blood pressure, mmHg (reciprocal)	744×10^{-9}	272×10^{-9}	31×10^{-9}	10×10^{-9}	4×10^{-9}
Diastolic blood pressure, mmHg (untransformed)	129.299	27.3406	3.50096	1.35484	0.13123
Diastolic blood pressure, mmHg (log)	0.02346	0.00465	0.00063	0.00025	0.00002
Body mass index, kg/m^2 (untransformed)	16.5753	3.72427	0.08555	0.04632	0.04110
Body mass index, kg/m^2 (reciprocal)	0.00003	0.00001	2×10^{-7}	9×10^{-8}	8×10^{-8}
Waist circumference, cm (untransformed)	153.510	7.37484	2.84500	0.23556	0.39360
Waist circumference, cm (reciprocal of square root)	572×10^{-7}	20×10^{-7}	11×10^{-7}	1×10^{-7}	2×10^{-7}
Hip circumference, cm (untransformed)	67.9275	15.0183	1.83384	0.07978	0.26203
Hip circumference, cm (reciprocal)	$5{,}217 \times 10^{-10}$	$1{,}179 \times 10^{-10}$	157×10^{-10}	6×10^{-10}	24×10^{-10}
GHQ (untransformed)	5.67758	0.98590	0.01868	0.00478	0.00758
Units of alcohol/week (untransformed)	265.599	56.9350	3.01593	1.81940	1.06043
Mean number of cigarettes/day (untransformed)	60.0652	20.6478	1.97665	0.77751	1×10^{-20}

hence there were no random effects to be estimated at this level. Also for both serum ferritin and haemoglobin the between household variance component was zero. Table 2 summarises the coefficients of skewness and kurtosis for the individual, postcode sector and district health authority level random effects. The ordinary coefficient of kurtosis is used, for which the value is 3 for the normal distribution, rather than the coefficient of excess kurtosis. Significance levels indicating departures from normality are shown, based on tests described by D'Agostino, Belanger, and D'Agostino (1990), but histograms and normal plots were mainly used to assess non-normality. The individual level residuals for most of the untransformed variables showed marked non-normality and in all cases were more non-normal than those for the corresponding transformed variables. Of the untransformed variables the distributions of individual level residuals for serum ferritin, number of units of alcohol drunk per week, glycated haemoglobin and General Health Questionnaire score showed the greatest degree of non-normality.

The random effects at the postcode sector and district health authority levels generally did not differ markedly from normality. There were, however, clear departures from normality for untransformed serum ferritin (the normal plot for the postcode sector level residuals is shown in the upper panel of Figure 1) and units of alcohol drunk per week at both the postcode sector and district health authority levels. In addition the postcode sector level residuals were markedly non-normal for untransformed glycated haemoglobin, General Health Questionnaire score and mean number of cigarettes smoked per day. For all untransformed variables except diastolic blood pressure and waist circumference, the postcode sector level residuals showed greater departure from normality than the district health authority level residuals.

The transformations were fairly successful in terms of normalising the individual level residuals. None of the distributions of cluster level residuals for the transformed variables differed markedly from normality. When comparing the postcode sector level and district health authority level residuals for the transformed variables there was no clear tendency for the former to deviate further from normality than the latter. This contrasts with the results for the untransformed variables.

Evidence of the extent to which applying normalising transformations at the individual level may normalise cluster level random effects is provided by analyses of serum ferritin. Log transformation was successful in terms of normalising the highly skewed distribution of individual level residuals for this variable. The cluster level random effects differed more markedly from normality for serum ferritin in its untransformed state than for the other outcomes. In contrast, for the log transformation of this variable, the postcode sector and district health authority level random effects had skewness and kurtosis coefficients that suggest normality. The normal plot of the postcode sector level residuals for log serum ferritin is shown in the lower panel of Figure 1. This plot contrasts with the corresponding plot for untransformed serum ferritin.

The relationship between normality at the cluster and individual levels is illustrated in plots of the coefficients of skewness for the clustering levels against those at the individual level (Figure 2). These are drawn separately for the postcode sector and district health authority levels in the upper and lower panels, respectively. The graphs demonstrate the larger variance of the skewness coefficients for residuals at the individual level than at the cluster levels. There was an increasing relationship between skewness at the individual

Table 2. Coefficients of skewness and kurtosis of random effects distributions at the individual, postcode sector and district health authority (DHA) levels

Outcome	Individual level residuals			Postcode level residuals			DHA level residuals		
	obs	skewness	kurtosis	obs	skewness	kurtosis	obs	skewness	kurtosis
Serum cholesterol, mmol/litre (untransformed)	11,106	0.407‡	3.311‡	711	0.169	2.812	177	−0.016	2.672
Glycated haemoglobin, % (untransformed)	10,890	3.691‡	28.527‡	711	0.463‡	3.388	177	0.423*	3.061
Glycated haemoglobin, % (reciprocal)	10,890	−0.542‡	5.973‡	711	0.071	2.977	177	−0.053	2.905
Plasma fibrinogen, g/litre (untransformed)[1]	9,747	0.990‡	5.440‡	711	0.404‡	3.358			
Plasma fibrinogen, g/litre (square root)[1]	9,747	0.467‡	4.039‡	711	0.215*	3.174			
Serum ferritin, μg/litre (untransformed)	10,943	7.267‡	129.704‡	711	2.819‡	25.102‡	177	0.780‡	3.813*
Serum ferritin, μg/litre (log)	10,943	−0.241‡	3.412‡	711	−0.090	2.928	177	−0.037	2.715
Haemoglobin, g/dl (untransformed)	10,751	−0.286‡	3.840‡	711	−0.204*	3.025	177	−0.088	2.682
Haemoglobin, g/dl (square)	10,751	0.098‡	3.306‡	711	−0.064	2.933	177	−0.053	2.659
Systolic blood pressure, mmHg (untransformed)	12,556	0.932‡	4.418‡	711	0.173	2.971	177	−0.007	3.545
Systolic blood pressure, mmHg (reciprocal)	12,556	−0.102‡	3.017	711	0.071	2.965	177	0.178	3.410
Diastolic blood pressure, mmHg (untransformed)	12,556	0.488‡	3.830‡	711	0.032	3.291	177	0.341	3.394
Diastolic blood pressure, mmHg (log)	12,556	−0.094‡	3.306‡	711	−0.192*	3.698†	177	0.205	3.376
Body mass index,[2] (untransformed)	14,681	1.011‡	5.272‡	712	0.150	2.862	177	-0.052	2.742
Body mass index,[2] (reciprocal)	14,681	0.160‡	3.205‡	712	0.180*	2.721	177	0.049	2.934
Waist circumference, cm (untransformed)	13,297	0.433‡	3.068	711	0.066	3.379	177	0.198	3.199
Waist circumference, cm (reciprocal of square root)	13,297	0.103‡	2.566‡	711	0.116	3.369	177	−0.212	3.263
Hip circumference, cm (untransformed)	13,325	1.157‡	7.107‡	711	0.201*	3.169	177	0.165	3.244
Hip circumference, cm (reciprocal)	13,325	−0.258‡	4.094‡	711	0.067	2.904	177	−0.065	2.925
GHQ (untransformed)	15,335	2.056‡	6.920‡	712	0.478‡	2.984	177	0.361*	2.915
Units of alcohol/week (untransformed)	15,792	4.190‡	41.233‡	712	1.192‡	5.541‡	177	1.019‡	4.778†
Mean number of cigarettes/day (untransformed)	4,327	1.155‡	7.028‡	707	0.573‡	3.714†	177	0.354	3.718

*$p < 0.05$; † $p < 0.01$; ‡ $p < 0.001$

[1] The between district health authority variance component for fibrinogen was 0

[2] Weight (kg)/height $(m)^2$

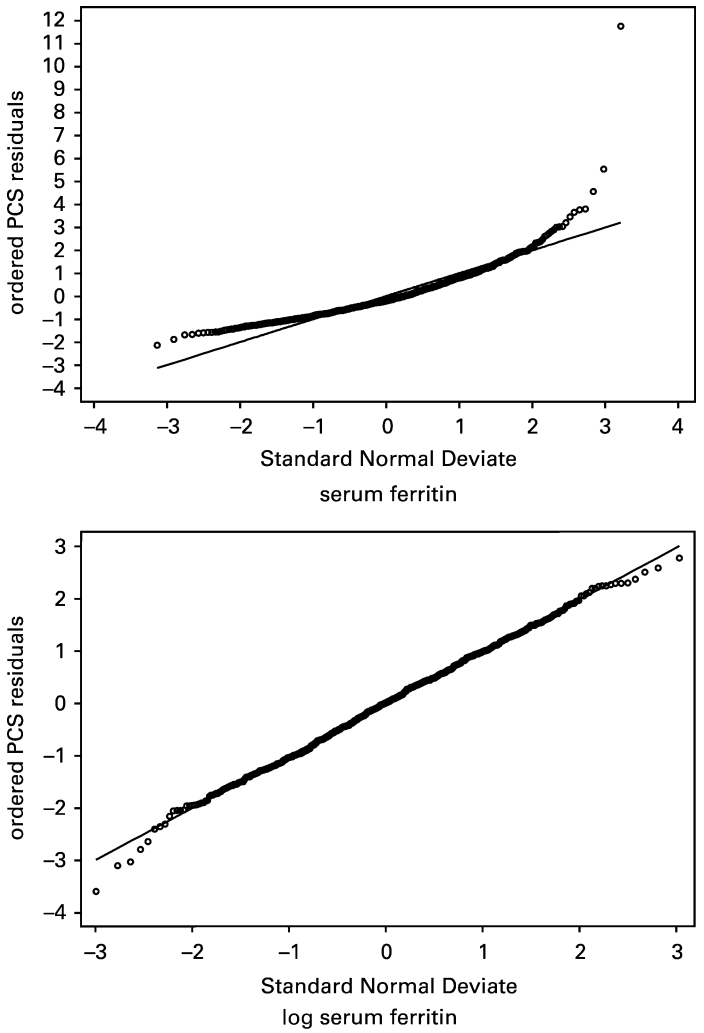

Fig. 1. Normal probability plot of the postcode sector (PCS) residuals for serum ferritin and log serum ferritin

level and skewness at both clustering levels. There was no clear relationship for kurtosis between the individual and clustering levels although, again, there was greater variance amongst the kurtosis coefficients for residuals at the individual level.

4. Discussion

The assumption of normality of cluster level random effects is generally difficult to assess in multilevel models as, often, insufficient numbers of clusters are available in studies to describe the distribution. The availability of data from large-scale complex surveys provides an opportunity to test the validity of the assumption for some common outcomes. This study used data from the Health Survey for England 1994 to obtain empirical distributions of cluster level random effects. These random effects represent the distribution at each level after the variation between clusters at higher levels has been

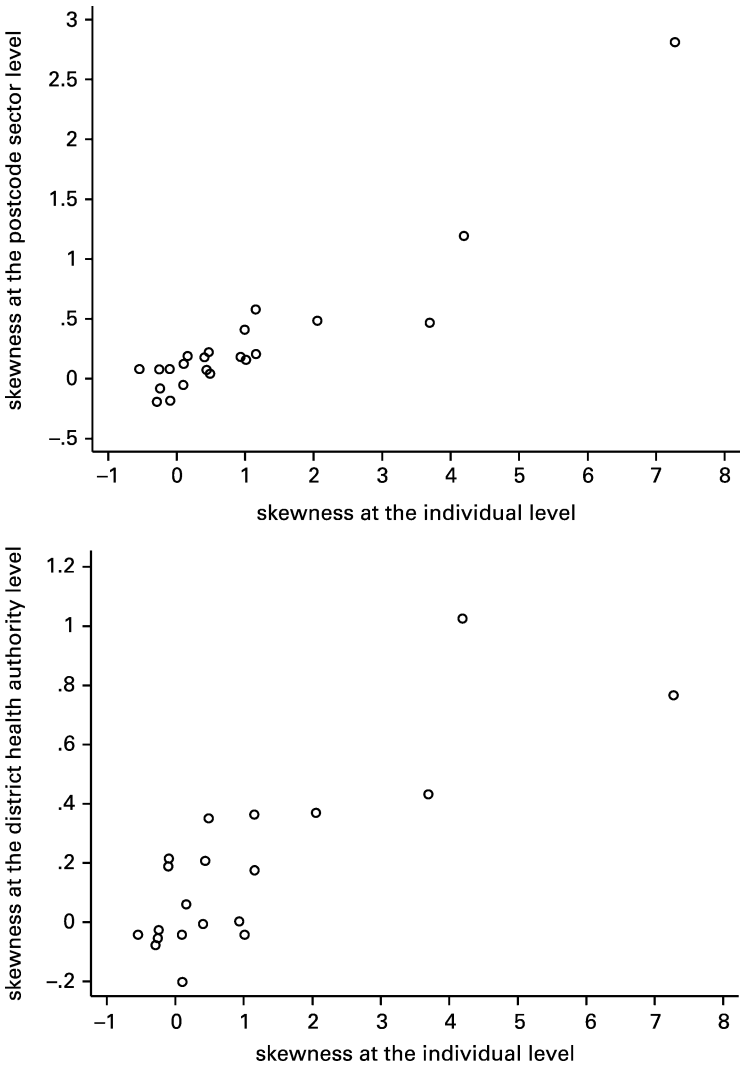

Fig. 2. Coefficients of skewness for the postcode sector and district health authority level residuals versus individual level residuals

accounted for and, therefore, represent the distribution that might be expected if sampling took place at a given level, stratified by clusters at higher levels. The Health Survey for England 1994 dataset contained large numbers of clusters at the postcode sector and district health authority levels. The structure of the UK National Health Service has undergone changes since the survey was carried out. There are now 28 strategic health authorities responsible for monitoring performance and around 400 primary care trusts responsible for the purchase and provision of health care and assessment of health care needs. The findings of this study are still of relevance to the analysis of health data in England as these changes are merely administrative and do not reflect changes in the composition of the clusters. The findings are also relevant to complex health surveys in other countries that are based on the sampling of organisational or administrative units.

Postcode sectors are similar in population size to electoral wards and the district health authorities of the 1994 Health Survey for England were based on towns and small counties.

The findings of this study suggest that normality of cluster level effects is not an unreasonable assumption to make for many continuous health outcomes. The random effects of a few variables, however, did show marked positive skewness at the clustering levels. It is possible that non-normality in these variables may be explained by cluster specific factors. Unfortunately such data were not available in the Health Survey for England 1994 to test this hypothesis. There was evidence of a relationship between the degree of non-normality at the individual level and the cluster levels. Only those variables with significant skewness or kurtosis at the individual level showed marked non-normality at the cluster level. There was also an increasing relationship between skewness at the individual level and at both clustering levels. It is not straightforward to make a reliable prediction of non-normality at the clustering levels on the basis of the individual level distribution, but marked non-normality at the individual level may be a useful marker. For most outcomes, transformations that normalised the individual level distributions also achieved a better approximation to normality at the postcode sector and district health authority levels. This was most noticeable for serum ferritin, which untransformed was the most skewed. An implication of the results of this study is that for data where distributional assumptions are questionable, transformations appropriate for normalising the individual level residuals may also normalise the cluster level random effects. Although only nine power transformations were tested in this study for the main analyses, they were sufficient to normalise almost all distributions, thus answering the question of whether normalising the individual level residuals also normalised the cluster level random effects.

The cluster level distributions of the outcomes were investigated using shrunken random effects. This was done so that the random effects estimated using information from smaller clusters were down-weighted to reflect their imprecision. Lange and Ryan (1989) presented an alternative approach for overcoming the problem of random effects having different variances that entails using weighted normal plots to assess the extent of deviation from normality. The weighting method used is identical to that used for shrinkage in multilevel models. The problem with both approaches is that outlying clusters with smaller numbers of subjects are less likely to be seen as outliers after shrinkage (Langford and Lewis 1998). For example if all the outlying clusters happen to contain relatively few subjects, then shrinkage may give a misleading impression of similarity (Duncan, Jones, and Moon 1998). This may have the consequence that the degree of non-normality at the cluster level is underestimated since the outliers will be pulled towards the overall mean. The problem of how best to accommodate the need to allow for the varying precision with which the cluster level random effects are estimated whilst obtaining an accurate estimate of their distribution has yet to be solved (Marshall and Spiegelhalter 2001).

The sizes of the variance components for the untransformed variables in this study generally differ slightly from those calculated by Gulliford, Ukoumunne, and Chinn (1999) for the same outcomes using the PROC VARCOMP procedure within SAS software (SAS Institute 1990). The RIGLS estimation procedure used by the MLwiN

software in this study is identical to the restricted maximum likelihood estimation procedure where the assumption of normality is used whereas the PROC VARCOMP procedure obtains moment based estimates. The differences observed between the estimation procedures are due to the unbalanced design of the Health Survey for England 1994 (Harville 1977). As the differences between the estimated variance components are generally not large, we do not believe the use of a different estimation method materially influenced the main findings of this study.

This study has focussed on the analysis of continuous health outcomes. Dichotomous outcomes are common in public health and health services research studies. The multilevel model extension to logistic regression requires the assumption that the distribution of cluster specific log odds is normal. Further research is needed to investigate whether this assumption holds for dichotomous health outcomes and to establish the degree of non-normality beyond which inferences made from multilevel models are invalid for all outcomes generally.

5. References

Colhoun, H. and Prescott-Clarke, P. (1996). Health Survey for England 1994. London: HMSO.

D'Agostino, R.B., Belanger, A., and D'Agostino, R.B. Jr. (1990). A Suggestion for Using Powerful and Informative Tests of Normality. The American Statistician, 44, 316–321.

Donaldson, R.J. and Donaldson, L.J. (1993). Essential Public Health Medicine. Plymouth: Petroc Press.

Donner, A. and Klar, N. (2000). Design and Analysis of Cluster Randomization Trials in Health Research. London: Arnold.

Duncan, C., Jones, K., and Moon, G. (1998). Context, Composition and Heterogeneity: Using Multilevel Models in Health Research. Social Science and Medicine, 46, 97–117.

Feng, Z., Diehr, P., Peterson, A., and McLerran, D. (2001). Selected Statistical Issues in Group Randomized Trials. Annual Review of Public Health, 22, 167–187.

Goldberg, D.P. (1972). The Detection of Psychiatric Illness by Questionnaire. London: Oxford University Press.

Goldstein, H. (1995). Multilevel Statistical Models (2nd ed.) London: Edward Arnold.

Gulliford, M.C., Ukoumunne, O.C., and Chinn, S. (1999). Components of Variance and Intraclass Correlations for the Design of Community-based Surveys and Intervention Studies: Data from the Health Survey for England 1994. American Journal of Epidemiology, 149, 876–883.

Harville, D.A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems (with Comment). Journal of the American Statistical Association, 72, 320–340.

Lange, N. and Ryan, I. (1989). Assessing Normality in Random Effects Models. Annals of Statistics, 17, 624–642.

Langford, I.H. and Lewis, T. (1998). Outliers in Multilevel Data (with Discussion). Journal of the Royal Statistical Society, Series A, 161, 121–160.

Marshall, E.C. and Spiegelhalter, D.J. (2001). Institutional Performance. In Multilevel Modelling of Health Statistics, A.H. Leyland and H. Goldstein (eds). Chichester: Wiley.

Moser, C.A. and Kalton, G. (1971). Survey Methods in Social Investigations (2nd ed.) London: Heinemann.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I., and Lewis, T. (2000). A User's Guide to MLwiN. London: Institute of Education.

Rice, N. and Leyland, A. (1996). Multilevel Models: Applications to Health Data. Journal of Health Services Research and Policy, 1, 154–164.

Royston, P. (1991). Comment on sg3.4 and an Improved D'Agostino Test. Stata Technical Bulletin, 3, 23–24.

SAS Institute, Inc. (1990). SAS/STAT User's Guide, Version 6 (4th ed.), Cary, NC: SAS Institute, Inc., 2, 1127–1134.

Solomon, P.J. (1985). Transformations for Components of Variance and Covariance. Biometrika, 72, 233–239.

StataCorp, (2000). Stata Statistical Software: Release 7.0. College Station, TX: Stata Corporation.

Turner, R.M., Omar, R.Z., and Thompson, S.G. (2001). Bayesian Methods of Analysis for Cluster Randomized Trials with Binary Outcome Data. Statistics in Medicine, 20 453–472.

Received March 2003
Revised December 2003