Multivariate Behavioral Research, *36* (2), 279-298

Multilevel Modeling of Cognitive Function in Schizophrenic Patients and Their First Degree Relatives

Sophia Rabe-Hesketh
Department of Biostatistics and Computing, Institute of Psychiatry, London

Timothea Toulopoulou and Robin M. Murray
Department of Psychological Medicine, Institute of Psychiatry, London

We describe multilevel modeling of cognitive function in subjects with schizophrenia, their healthy first degree relatives and controls. The purpose of the study was to compare mean cognitive performance between the three groups after adjusting for various covariates, as well as to investigate differences in the variances. Multilevel models were required because subjects were nested within families and some of the measures were repeated several times on the same subject. The following four methodological issues that arose during the analysis of the data are discussed. First, when the random effects distribution was not normal, non-parametric maximum likelihood (NPML) was employed, leading to a different conclusion than the conventional multilevel model regarding one of the main study hypotheses. Second, the between-subject (within-family) variance was allowed to differ between the three groups. This corresponded to the variance at level 1 or level 2 depending on whether repeated measures were analyzed. Third, a positively skewed response was analyzed using a number of different generalized linear mixed models. Finally, penalized quasilikelihood (PQL) estimates for a binomial response were compared with estimates obtained using Gaussian quadrature. A small simulation study was carried out to assess the accuracy of the latter.

Introduction

We briefly describe the background and design of the study. Subsequent sections discuss four methodological issues encountered in the analysis of the data.

Despite strong evidence for the involvement of genes in susceptibility to schizophrenia, the nature of these genes and the traits they transmit are as yet

---

This work was partly funded by a Stanley Foundation grant given to Professor Robin Murray. Timothea Toulopoulou was awarded a Scholarship by the Department of Psychological Medicine at the Institute of Psychiatry. She is indebted to Dr Robin Morris for supervising the neuropsychological aspects of the study. We would also like especially to thank Darren Mockler, Heather King, Tonmoy Sharma, Sophia Frangou, Thordur Sigmundsson, Anton Grech, Jessica Yakeley and Brian Everitt for their invaluable contributions.

unknown (Faraone, Seidman, Kremem, Pepple, Lyons & Tsuang, 1995; Kendler, McGuire, Gruenberg, O'Hara, Spellman & Walsh, 1993). Unraveling this complex and elusive inheritance mechanism may necessitate the employment of indirect measures that will eventually help to determine the schizophrenia genotype by firstly defining more precisely its phenotype (McGuffin, Asherton, Owen & Farmer, 1994; Faraone et al., 1995, 1996). The approach assumes that if genetic susceptibility is reflected in phenotypic variability, then exploration of the disorder's phenotypic boundaries might lead to better characterization of its genetic constitution. Several lines of enquiry have suggested that genetic predisposition to schizophrenia can also be expressed as liability to non-specific, non-psychotic dysfunctions including cognitive impairment (Saykin et al., 1991, 1994; Kremen, Seidman, Pepple, Lyons, Tsuang & Faraone, 1994). This position has been further reinforced by the emergence of evidence suggesting an increased prevalence of deficits of a similar nature to those observed in schizophrenic patients in their otherwise healthy biological relatives (Toulopoulou, Morris, Rabe-Hesketh, King & Murray, 1999; Kremen et al., 1994).

The present study set out to explore whether specific impaired cognitive functions are risk indicators for schizophrenia. We hypothesized that (a) cognitive impairment is present in schizophrenia (b) cognitive impairment is present in some of the first-degree relatives of schizophrenic patients (c) since not all relatives are expected to be carriers of the schizophrenia genotype, a greater dispersion of scores (i.e., greater variance) should be seen in the relative sample when compared to controls.

Seventy schizophrenic patients and 115 of their healthy first-degree relatives from 59 families and 66 normal unrelated controls underwent a series of neuropsychological examinations assessing intelligence, verbal and visual episodic memory, spatial working memory, shifting mental sets and planning ability. There were therefore a total of 251 subjects from 125 families, although not all measures were available on all subjects. Cases with missing values on variables included in a model were omitted from the analysis. When data are not missing at random, this can lead to biases in the parameter estimates and a better method would have been to use, for example, multiple imputation.

Subject recruitment has been described in some detail elsewhere (see e.g., Griffiths et al., 1998). Briefly, families with one or more schizophrenic members (according to DSM-III-R criteria) were acquired as part of the Maudsley Family Study by referrals through a network of psychiatric clinics and voluntary care organizations across the United Kingdom. The first 59 families who were referred to us and who agreed to undergo a series of neuropsychological assessments are included in the present sample. Control

subjects with no personal or family history, up to second degree, of psychotic illness were acquired from a pool of controls obtained for previous studies conducted in the Institute of Psychiatry, from members of staff at the Bethlem and Maudsley Trust and via advertisements in the local press. The control group may not be entirely representative so that differences in the means and variances between the control and patient groups cannot necessarily be attributed to the schizophrenia phenotype. The results of all analyses must therefore be interpreted cautiously.

Linear Regression with Random Effects for Families

Analysis of variance (ANOVA) or linear regression could be carried out in order to address the hypothesis that cognitive abnormalities are present in schizophrenic subjects and to a lesser degree in their healthy relatives. However, subjects within the same family are expected to perform more similarly than subjects from different families in the various cognitive tasks owing to shared genes and shared environments. The observations may not therefore be assumed to be independent as required for inference in ANOVA or simple regression. One way of correcting for the clustering of observations is to use the sandwich variance estimator (also known as the robust variance estimator) for clustered data (Diggle, Liang, & Zegar, 1996, p.68-69). However, we prefer to model the interdependence between members of the same family directly, because the interdependence is of interest in its own right. We can model the interdependence by introducing a random effect for families into the linear regression model. For example, the IQ score (WAIS-R) may be modeled a

$$(1) \qquad Y_{ij} = \beta_0 + \beta_1 x_{Aij} + \beta_2 x_{Fij} + \beta_3 x_{Pij} + \beta_4 x_{Rij} + u_{oi} + r_{ij}$$

where x_A is the age, centered around the mean age of 42.3, x_F is an indicator variable for females and x_P and x_R are indicator variables for the patient group and relative group respectively. The indices i and j denote families and subjects respectively, u_{oi} is the random effect for family i and r_{ij} is the residual for subject j in family i. The random effect for family may be interpreted as the effect of any family-specific predictors that have not been controlled for (or even measured). These predictors may include shared genetic or environmental factors. Similarly, the residual term for subjects within families may be interpreted as the effect of characteristics specific to the individual, plus measurement error. This model is a hierarchical or multilevel model because subjects, the "level 1 units" are nested within families, the "level 2 units". The random terms u_{oi} and r_{ij} are assumed to be

independently distributed with zero means and constant variances, $\sigma^2(u_0)$ and $\sigma^2(r)$, respectively. (Later we will allow the level 1 variance to differ between groups.) Further, the random terms are assumed to be uncorrelated with each other and with the covariates. These assumptions imply a compound symmetric structure of the residual covariance matrix within families, with all family members having the same variance $\sigma^2(u_0) + \sigma^2(r)$, and all pairs of members of the same family having the same correlation $\rho = \sigma^2(u_0) / [\sigma^2(u_0) + \sigma^2(r)]$, called the intraclass correlation (ICC), conditional on the covariates. The intraclass correlation may be interpreted as the proportion of the residual variance that is due to genetic and environmental factors that are shared among family members. However, separation of variance components into genetic, shared environment and unique environment is possible only if family members with different degrees of relatedness are available, for example monozygotic and dizygotic twins (see e.g., Sham, 1998), and will not be pursued further in this article.

The random effects u_{0i} and r_{ij} are usually assumed to have a normal distribution. In this case, estimation of the model by maximum likelihood is straightforward because the likelihood, given by

$$\ell = \prod_i \int \prod_j P(Y_{ij}|u_{0i}, \mathbf{x}_{ij}) P(u_{0i}) du_{0i} \tag{2}$$

has a closed form. The model may be fitted using most standard statistical packages, for example Stata, S-Plus or SAS, as well as the multilevel modeling package MlwiN (Goldstein et al. 1998). The latter was used to obtain the parameter estimates (based on 118 families and 236 subjects) shown in Table 1. We present the standard deviations of random terms because they have the same scale as the other parameters and are therefore easier to interpret than variances. The delta-method was used to estimate the standard error of the between-family standard deviation from the standard error of the variance given in MLwiN. The mean IQ score for men aged 42.3 in the control group is estimated as 112.3. The mean IQ score in the patient group is over 15.3 units lower than that of the control group ($p < 0.001$) and the relatives differ from the controls by only about 2.1 units (n.s.), after controlling for sex and age. The intraclass correlation is estimated as 0.50, that is, half the variance in IQ scores remaining after controlling for group, age and sex, is shared between members of the same family.

The parameter estimates for the linear random intercept model in Equation 1 are a weighted average of the within-family estimates and between-family estimates (see Hsiao, 1986, p. 36). For example, we can obtain both within and between-family estimates of the difference in mean IQ between the patient and relative groups. The within-family estimate, also known as a fixed effects

Table 1
Parameter Estimates for the Model in Equation 1 with Normal and Non-Parametric Random Effects Distributions

	Normal Random Effect			NPML		
Parameter	Estimate	se	*p*	Estimate	se	*p*
Intercept	112.29	2.39		114.23	2.28	
Age	0.14	0.069	0.04	0.16	0.067	0.02
Female	-1.21	1.87	0.52	-1.64	1.71	0.34
Patient	-15.32	3.10	<0.001	-18.89	2.87	<0.001
Relative	-2.12	3.13	0.50	-6.21	3.00	0.04
$\sigma(u_0)$	11.98	1.29		12.62	-	
$\sigma(r)$	11.90	0.74		11.36	0.65	
ICC	0.50	0.07		0.55		
Log-likelihood	-978.927			-971.685		

or conditional estimate, is a weighted mean of the within-family mean differences between patients and their relatives. The between-family estimate is the slope of the regression of the families' mean IQ scores on the proportion of family members (contributing to the families' mean scores) who are patients, that is, the expected difference in mean IQ score between a family represented wholly by its affected family members and a family represented wholly by its unaffected family members. Both within and between family estimates were obtained for the model in Equation 1, without the random effect for family and with the dummy variable for relatives replaced by a dummy variable for controls. The within-family estimate (standard error) of the difference in mean between patients and relatives was 12.84 (2.63). The between-family estimate was 17.82 (7.76) and the random effects estimate was 13.20 (2.36). The between family estimate is the least precise because individual differences between families have not been controlled for and therefore contribute to the residual error variance of the regression. The between-family estimate is also more vulnerable to bias. In contrast, the within-family estimate eliminates differences between families completely. The random effects estimate lies between the between and within-family estimates and has the smallest standard error because both sources of information have been combined. However, this is achieved only by assuming that the families are a random sample from a population of families and that the random effect for families has a normal distribution

We can assess this distributional assumption by estimating or predicting the values of u_{0i} by their conditional expectations given the observed data. The assumption that r_{ij} in Equation 1 has a normal distribution with mean 0 and variance $\sigma^2(r)$ specifies the conditional density of the responses given the (vector of) explanatory variables $\mathbf{x}_{ij}$ and the random effect u_{oi} as

$$P(Y_{ij} \mid u_{0i}, \mathbf{x}_{ij}) = \mathrm{N}[(\mathbf{x}_{ij}^T\boldsymbol{\beta} + u_{0i}, \sigma^2(r)] \tag{3}$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients. The posterior density of the random effects given the observed responses and explanatory variables, $P[u_{0i} \mid \{Y_{ij}\}, \mathbf{x}_{ij}]$, may be obtained by applying Bayes Theorem and the random effects are estimated by their conditional expectations

$$E[u_{0i}|\{Y_{ij}\},\mathbf{x}_{ij}] = \int u_{0i} P[u_{0i}|\{Y_{ij}\},\mathbf{x}_{ij}]du_{0i} \tag{4}$$

known as *empirical Bayes* estimates (see for example Morris, 1983). A histogram of the empirical Bayes estimates of the random effects also interpretable as level 2 residuals for model 1 is given in Figure 1 (bottom left panel) and a normal Q-Q plot is shown in Figure 2. The distribution appears to be slightly non-normal.

Non-Parametric Maximum Likelihood

We will now investigate how much the parameter estimates change when we do not assume a normal random effects distribution and estimate the model by non-parametric maximum likelihood (NPML) (see e.g., Simar, 1976; Laird, 1978; Lindsay, Clogg & Grego, 1991 or Aitkin,1999). The NPML estimate of the random effects distribution is a discrete distribution on a finite number of mass-points with locations z_k and weights π_k so that the likelihood in Equation 2 becomes

$$\ell = \prod_i \sum_k \pi_k \prod_j P(Y_{ij}|u_{0i} = z_k, \mathbf{x}_{ij}) \tag{5}$$

This model is also known as a semi-parametric mixture model. Some small simulation studies indicate that NPML estimates are approximately unbiased when the true random effects distribution is normal (Davies, 1987; Follman & Lambert, 1989; Rabe-Hesketh & Pickles, 2001).

A program called gllamm (Rabe-Hesketh & Pickles, 1999; Rabe-Hesketh et al., 2000, 2001), written in Stata, was used to maximize the likelihood for a given number of mass-points. (gllamm can be downloaded from http://www.iop.kcl.ac.uk/iop/departments/biocomp/programs/gllamm.html.) The

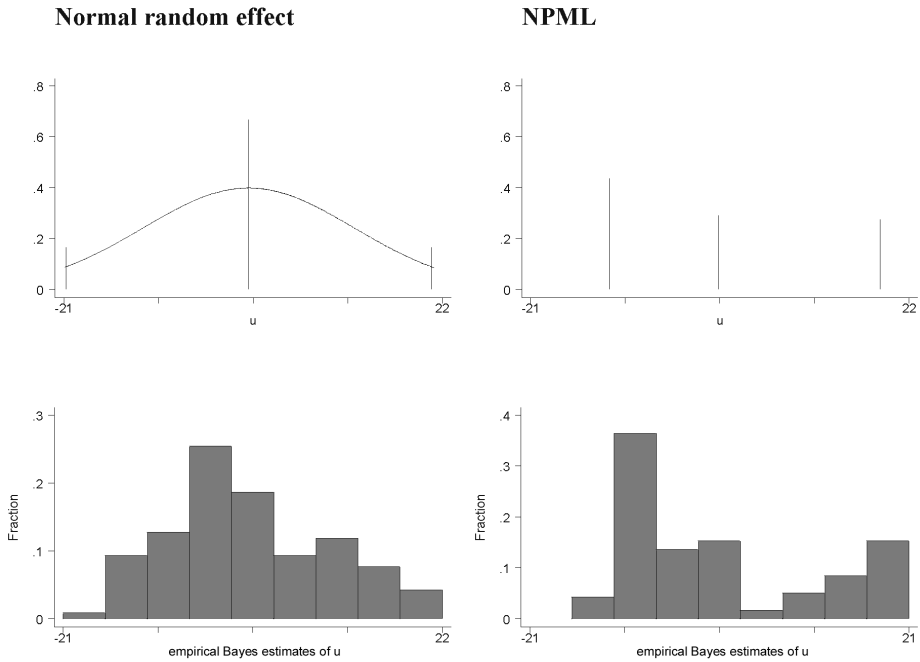

Figure 1
Random Effects Distributions and Histograms of Empirical Bayes Estimates of the Random Effects for the Model in Equation 1 with Normal and Non-Parametric Random Effects Distributions

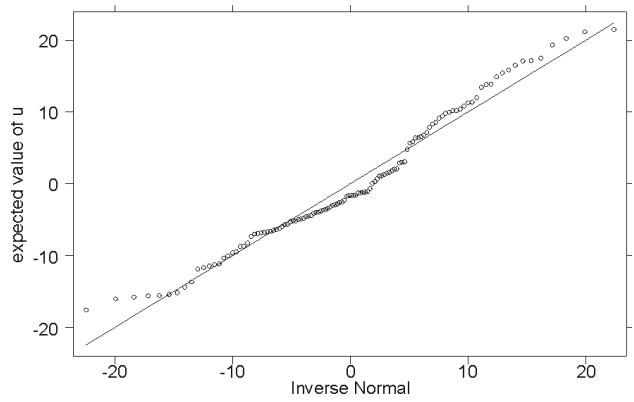

Figure 2
Q-Q Plot of the Empirical Bayes Estimates of Random Effects Assuming a Normal Random Effects Distribution

program uses a modified version of the Newton-Raphson algorithm implemented in Stata's ml functions to maximize the log-likelihood and provides standard errors based on the numerically estimated Hessian matrix. In order to obtain the NPML solution, we need to identify the maximum number of mass-points beyond which the likelihood does not increase any further. This was done using the Gateaux derivative method (Heckman & Singer,1984; Follmann & Lambert,1989 and Davies & Pickles, 1987). Starting with two masses, a new mass with very low probability was moved across a range of locations (-50 to 50 in 1000 equal steps) keeping all other parameter estimates and mass-points fixed. If the likelihood increased at any location, a further mass-point was introduced. The method was repeated until no further increase in likelihood could be achieved. Two mass-points gave a log-likelihood of -974.90, three mass-points gave a log-likelihood of -971.68 and no further mass-points were required. The mass-points, centered around the mean, were located at (-12.0, 0.38, 18.7) with probabilities (0.44, 0.29, 0.27) giving a standard deviation of 12.62. The semi-parametric mixture model has three extra parameters compared with the parametric model because two location and two probability parameters are required to specify the random effects distribution (the third being determined by setting the sum of the probabilities to one and the mean location to zero) instead of one standard deviation parameter required for the parametric model. The associated increase in twice the log-likelihood of 14.4 suggests a better fit of the NPML solution although the likelihood ratio test is not valid here because the models are not nested. The parameter estimates for this model are shown in Table 1 (second column). The NPML solution attributes a greater portion of the variance to the random intercept term, a result that is consistent with the findings of Rabe-Hesketh and Pickles (2001). The estimated adjusted mean difference between relatives and controls has increased substantially and is now significant at the 5% level. Assuming a normal random effects distribution has therefore changed our conclusion regarding hypothesis b.

By substituting specific locations and weights, Equation 5 can also be used to approximate the integral in Equation 2 numerically when a normal random effects distribution is assumed. These locations and weights, given by Gaussian quadrature (Bock & Lieberman, 1972, Aitkin, 1999), may be compared with the masses estimated by NPML to informally assess how "non-normal" the random effects are. The three quadrature points and normal density corresponding to the parameters estimated using MLwiN are plotted in Figure 1 (upper left panel), next to the mass-points of the semiparametric mixture model (upper right panel). The latter is asymmetric and differs substantially from the former. The corresponding frequency distributions of the empirical Bayes estimates of the random effects for the two models are shown in the second row of Figure 1. Here it is clear that

for the semiparametric mixture model, there are more level-2 residuals at locations near the outer mass-points than under the normal random effects assumption. Extremely intelligent or unintelligent families are therefore more easily accommodated in the random part of the model so that they may contribute less to the fixed effects estimates. This may result in an increased contribution of the within-family differences to the estimates with the difference in mean IQ between patients and relatives now estimated as 12.68 (2.30) and with coefficients of age and sex very close to the conditional estimates. In the case of the Rasch model, Lindsay et al. (1991) showed that parameter estimates for the semiparametric mixture model are identical to those obtained using conditional maximum likelihood.) Using the NPML solution, we can therefore conclude that the mean (sex and age-adjusted) IQ score of relatives lies between the mean scores of schizophrenics and controls and differs significantly from both of them.

Heteroskedasticity at Level 1

In order to address hypothesis (c), that relatives show greater variability (within families) than controls, model (1) was again estimated, this time allowing the level 1 variance to differ between the three groups. Assuming normally distributed random effects (using MLwiN), the estimated within-family standard deviations were 13.14, 11.34 and 9.20 for patients, relatives and controls respectively. The change in deviance associated with estimating two further parameters was 1.58 so that there was no evidence for differences in the variances. Allowing the variances to differ for the three mass-point solution, and re-fitting the mass-point locations and masses, gave standard deviation estimates of 13.30, 10.70 and 8.11 for patients, relatives and controls respectively. The deviance decreased by 5.88 ($DF = 2, p = 0.05$). The parameter estimates for this model are shown in Table 2. Testing the difference between the variance estimates (parameterized as log standard deviations) for relatives and controls using a Wald test, gave $p = 0.08$ whereas the difference between the variances of the schizophrenic and control groups was highly significant ($p = 0.009$).

Three-Level Models

More complex models are required to analyze measures taken several times on each subject under different conditions. One example is the Tower of London task (Morris et al. 1993). Briefly, in the Tower of London task, two arrangements of three colored disks slotted onto three rods of unequal length are presented on a computer screen. The top half of the screen

Table 2
Semiparametric Mixture Model Allowing the Level 1 Variance to Differ Between Groups

	NPML		
Parameter	Estimate	se	*p*
Intercept	114.94	2.02	
Age	0.17	0.064	0.02
Female	-1.32	1.62	0.34
Patient	-20.40	2.67	<0.001
Relative	-8.48	2.70	0.04
$\sigma(u_0)$	13.57	-	
$\sigma(r_C)$	8.11	1.16	
$\sigma(r_R)$	10.70	0.96	
$\sigma(r_P)$	13.30	1.48	
Log-likelihood	-968.743		

displays the goal arrangement, while the lower half of the screen displays the starting position. The subject has to rearrange the bottom array to attain the goal arrangement according to specified rules, making the fewest possible moves. The level of difficulty is increased by increasing the minimum number of moves required. Measures of interest include the time taken to complete the tower and whether this is achieved using the minimum number of moves. For each level of difficulty, the subject is also asked to simply copy the moves demonstrated on the top half of the screen. This control task provides a measure of motor response latency.

Since each subject completed each of the three levels of difficulty (excluding the control task), all three measures may be modeled together using a 'repeated measures' model that includes a further random effect for subjects, giving a 3-level model. The time to complete the Tower of London was log transformed because it had a positively skewed distribution. The log completion time of subject *j* in family *i* on occasion *k* was modeled as

$$\log(Y_{ijk}) \mid \mathbf{x}_{ijk},\ u_{0ij}^{(2)},\ u_{0i}^{(3)} \sim \mathrm{N}(\mu_{ijk},\ \sigma^2) \tag{6}$$

where

$$\mu_{ijk} = \beta_0 + \beta_1 x_{Aij} + \beta_2 x_{Fij} + \beta_3 x_{Pij} + \beta_4 x_{Rij} + \beta_5 x_{Lijk} + \beta_5 x_{Cijk} + u_{oij}^{(2)} + u_{oi}^{(3)} \tag{7}$$

and $u_{oij}^{(2)}$ and $u_{oi}^{(3)}$ are the level 2 (subjects) and level 3 (families) random effects, respectively. These random effects are assumed to be independent of each other. Two new, time-varying, covariates are used here, x_{Lijk} and $x_{Cijk,}$ the level of difficulty (coded as -1, 0, 1) and the log time taken to complete the control task, respectively. This model assumes that the intraclass correlations among repeated measures are constant, that is, the residual covariance matrix follows compound symmetry, an assumption that may not hold if there is for example serial autocorrelation (Diggle, 1988). Since we only have 3 repeated measures per subject, too few to provide reliable estimates for the autocorrelation, we will ignore this issue here.

The analysis was based on 212 subjects from 112 familes who attempted the Tower of London; all of these subjects had observations on all three occasions. The model was estimated using gllamm by numerically integrating the joint likelihood of the data and the random effects over the random effects distributions using 16-point Gaussian quadrature. The resulting parameter estimates for this model are shown in Table 3 (first column). Using eight quadrature points did not change the parameter estimates after rounding to two significant figures, suggesting that the 16 quadrature points were adequate. The parameters indicate that the completion times increase significantly with age and are higher for females and for patients. The relatives are slower than the controls and quicker than the patients, but these differences are not significant.

Heteroskedasticity at Level 2

To test hypothesis (c), we again allowed the variance component for subjects to differ between the three groups. This time the between-subject variance is at level 2 and this model was fitted using gllamm, again employing 16-point Gaussian quadrature. The standard deviations were estimated as 0.00, 0.22 and 0.00 for patients, relatives and controls respectively and the change in deviance was 3.42 ($DF = 2$, $p = 0.18$). The difference between relatives and controls was not significant using the Wald test ($p = 0.38$) but the difference between relatives and patients was nearly significant at the 5% level ($p = 0.08$). Again, the same parameter estimates were obtained using 8 point quadrature.

Multilevel Generalized Linear Mixed Models (GLMMS)

Instead of log-transforming the time taken to complete the Tower of London task, we could model the time on its natural scale, using a generalized linear model. A natural choice of error distribution for a positively skewed response would be the gamma distribution and we will combine this with a

Table 3
Parameter Estimates for (a) a Three Level Model for the Log Time to Completion of the Tower of London Assuming an Identity Link and Normal Errors and (b) a Three Level Model for the Time to Completion of the Tower of London Assuming a Log Link and Gamma Errors

	lognormal time		gamma time	
Parameter	Estimate	se	Estimate	se
Intercept	1.27	0.27	1.47	0.20
Age	0.0081	0.0023	0.0082	0.0026
Female	0.14	0.056	0.16	0.064
Patient	0.19	0.083	0.21	0.093
Relative	0.10	0.077	0.11	0.087
Level	0.42	0.036	0.42	0.039
C. task	0.70	0.10	0.67	0.12
$\sigma[u_0^{(3)}]$	0.21	0.044	0.25	0.049
$\sigma[u_0^{(2)}]$	0.076	0.13	0.23	0.048
σ	0.63	0.021	0.59	0.019

log link. As for the linear regression model with random effects, the random effect may be added to the linear predictor giving

$$Y_{ijk} \mid \mathbf{x}_{ijk}, u_{0ij}^{(2)}, u_{0i}^{(3)} \sim \text{gamma}(\mu_{ijk}, \sigma^2), \tag{8}$$

and

$$\log(\mu_{ijk}) = \beta_0 + \beta_1 x_{Aij} + \beta_2 x_{Fij} + \beta_3 x_{Pij} + \beta_4 x_{Rij} + \beta_5 x_{Lijk} + \beta_5 x_{Cijk} + u_{0ij}^{(2)} + u_{0i}^{(3)}. \tag{9}$$

where σ^2 is the squared coefficient of variation so that $\text{var}(Y_{ijk} \mid \mathbf{x}_{ijk}, u_{0ij}^{(2)}, u_{0i}^{(3)}) = \sigma^2 \mu_{ijk}^2$. Adding the random effects to the linear predictor in this way is appealing if one wishes to interpret the random effects as representing unmeasured covariates.

The linear model assumes a constant variance of the log completion time whereas the gamma model assumes a constant coefficient of variation, that is, the standard deviation in completion time is proportional to the mean. The two assumptions are approximately equivalent because the log of a variable with a gamma distribution and a small coefficient of variation σ has a constant variance of approximately σ^2 (McCullagh & Nelder, 1989, p. 285).

The log-normal model equates the mean of the log of Y to the linear predictor and the gamma model equates the log of the mean of Y to the linear predictor. In the gamma model, the exponentiated slope parameters are therefore ratios between mean completion times; this is more natural than the interpretation in the log-linear model where the exponentiated slope parameters are ratios between the *geometric* means of the completion times. However, the distinction may be academic, since, in the absence of random effects, linear regression of the log-transformed response gives consistent but possibly inefficient estimates of the regression parameters, except for the intercept, of the gamma model with log link [see McCullagh & Nelder (1989), p. 285-286.]

An important problem with generalized linear mixed models (GLMM) is that the likelihood in Equation 2 does not have a closed form solution except for certain combinations of distributions of the response and random effects. We will discuss this issue in more detail in the next section. Here we again used 16-point Gaussian quadrature to approximate the likelihood. The resulting parameter estimates are given in Table 3 (column 2). The fixed effects coefficients are remarkably similar to those for the model assuming normal errors of the log completion times. The standard deviation within subjects is about 0.59 times the mean, so that the coefficient of variation is close to the standard deviation of 0.63 of the log-normal model as expected. However, the level-2 standard deviation of the gamma model is considerably higher than that of the log-normal model. Allowing the level 2 variance to differ between groups gave standard deviation estimates of as 0.14, 0.28 and 0.16 for patients, relatives and controls, respectively and decreased the deviance by only 1.82 ($DF = 2$, $p = 0.40$).

Accuracy of Quasilikelihood and Quadrature

We used the following logistic mixed model to model the probability that subject j in family i completes the tower in the minimum number of moves on occasion k,

$$(10)\text{logit}(\pi_{ijk}) = \beta_0 + \beta_1 x_{Aij} + \beta_2 x_{Fij} + \beta_3 x_{Pij} + \beta_4 x_{Rij} + \beta_5 x_{Lijk} + u_{0ij}^{(2)} + u_{0i}^{(3)}.$$

Two methods that are frequently used to estimate generalized linear mixed model are *marginal quasilikelihood* (MQL) or *penalized quasilikelihood* (PQL) (Breslow & Clayton, 1993), methods implemented in MLwiN (Goldstein & Rasbash, 1996). In PQL, the inverse link as a function of the linear predictor including the random effects, for example, $p[\mathbf{x}_{ij}^T \beta + u_{0ij}^{(2)} + u_{0i}^{(3)}]$, is approximated by a first order Taylor expansion with

respect to the fixed effects (around the 'current' estimates) plus a first or second order Taylor expansion with respect to the random effects (around the 'current' empirical Bayes estimates of the random effects). This reduces the model to a linear model so that procedures for linear multilevel models may be used. MQL is less accurate than PQL because the random part is expanded about zero, not about the current empirical Bayes estimates of the random effects. The second order PQL parameter estimates for the model in Equation 10 (using data on 112 families, 212 subjects and 3 observations per subject) are shown in Table 4. Since the variance component for families was not significant, the estimates for the two-level model without this variance component are also shown.

We will now consider the accuracy of these estimates. McCulloch (1999) remarked "Unfortunately this penalized quasi-likelihood approach has worked poorly in practice. This has been especially the case where generalized linear mixed models are most needed, for example, with binary data." The method is particularly inaccurate when the binomial denominator and number of lower level units per higher level unit are small and when the variance components are large (see e.g., Breslow & Lin 1995; Rodíguez, G. & Goldman, 1995; Lin & Breslow, 1996). All three criteria appear to be met in the current example.

In order to evaluate the second order PQL estimates for the logistic regression model in Equation 10, we will first compare them with 30 point Gaussian quadrature estimates which are given in Table 4 for both the three-level and two-level models. We assessed the accuracy of the quadrature solution by investigating the change in parameter estimates when the number of quadrature points was increased from 6 to 12 to 20 to 30. For both models, the first increase resulted in changes in the parameter estimates in the second to fifth significant figure. The increases from 12 to 20 and from 20 to 30 resulted in changes in no more than the fourth significant figure (largest relative changes of 0.03%). Since the discrepancies between the PQL and quadrature estimates in Table 4 are greater for the two-level model, we will focus the rest of our discussion on this model. It is interesting to note that the estimated standard errors for PQL are lower than for quadrature, particularly for the between-subjects standard deviation. The same is true of the standard deviation estimate itself. The latter is likely to be at least partly due to a downward bias in the PQL estimate since this has been shown to be a particular problem for the level-2 standard deviation (see e.g. Lin & Breslow 1995) and since the estimate increased substantially from 0.82 for first order MQL to 1.21 for second order PQL

Goldstein et al. (1998) suggest using parametric bootstrap estimation to correct any bias in PQL parameter estimates for binomial data. The

Table 4
Parameter for Two- and Three Level Logistic Regression Models using PQL and Gaussian Quadrature

	Three level model				**Two level model**					
	2nd order PQL		30 quad. points		2nd order PQL		30 quad. points		Bootstrap	
	Estimate	se	Estimate	se	Estimate	se	Estimate	se	Estimate	se
Age	0.00049	0.011	0.00061	0.011	0.00073	0.010	0.00086	0.011	0.00076	0.011
Female	-0.31	0.31	-0.30	0.32	-0.31	0.29	-0.32	0.31	-0.32	0.33
Patient	-0.90	0.42	-0.88	0.43	-0.85	0.39	-0.87	0.42	-0.87	0.42
Relative	-0.050	0.38	-0.057	0.39	-0.019	0.35	-0.025	0.38	-0.021	0.51
Level	-1.75	0.18	-1.72	0.21	-1.65	0.17	-1.72	0.21	-1.70	0.19
Intercept	-1.57	0.33	-1.57	0.35	-1.48	0.31	-1.56	0.34	-1.53	0.34
$\sigma[u_0^{(3)}]$	0.45	0.40	0.47	0.60	-	-	-	-	-	-
$\sigma[u_0^{(2)}]$	1.26	0.22	1.28	0.34	1.21	0.17	1.36	0.28	1.36	0.28
LL			-279.53				-279.61			

parametric bootstrap procedure works as follows. The response vector is simulated n times from the initial PQL parameter values and the parameters are re-estimated for each simulation. This provides an estimate of the bias in the initial PQL parameter values that is used to correct these values. However, this single bias correction is unlikely to be sufficient because the bias depends on the 'true parameters' used to simulate the model. The procedure is therefore repeated by carrying out further sets of n simulations each, where each set uses the 'current' parameter values for the simulations and provides a bias estimate that is used to update these values for the next set. This procedure is repeated until convergence and the results of the last parameter update represent the final bias-corrected parameter estimates. Convergence is judged informally by inspecting trajectories of the bias-corrected parameter estimates both within and between sets of simulations. In order to be confident that the procedure has converged, we ran the bootstrap procedure in MLwiN with seven sets of n=3000 simulations each to obtain the parameter estimates shown in Table 4.

Simulation Study

The parametric boostrap estimates, particularly the between subject standard deviation estimate and its standard error, are much closer to the quadrature estimates than to the PQL estimates. In order to further evaluate the quadrature estimates, we carried out the following small simulation study. Using the same hierarchical structure and covariate values as the observed data, a new response variable was simulated 1000 times using the parameters estimated by 30 point quadrature (two-level model) shown in Table 4. Table 5 summarizes the results of estimating the model for each simulated dataset using 30 point Gaussian quadrature. Note that this is similar to carrying out a single set of the parametric bootstrap procedure and the bias estimate could be used to correct the quadrature parameter estimates.

The estimates of the intercept and of the coefficient of "level" have significant downward bias (at the 5% level) as does the standard deviation estimate of the random effect. The relative bias in these parameters is no more than about 2%, however, much smaller than the discrepancy in these parameters between the PQL and quadrature estimates. The small biases do not appear to be due to insufficient quadrature points since the parameter estimates of the first 200 simulations did not change when 60 quadrature points were used. The mean standard errors are close to the standard deviations over the1000 simulations and the coverage probabilities of the 95% confidence intervals do not differ significantly from 0.95 using an exact binomial test at the 5% level except for the standard deviation estimate where the coverage probability appears to be slightly

Table 5
Results of Estimating the Two-Level Logistic Regression Model on 1000 Simulated Dataset using 30 Point Gaussian Quadrature

	True value	Mean estimate	Standard deviation	Relative bias (%)	95% CI of relative bias (%)	Mean standard error	Coverage probability of 95% CI
Age	0.000860	0.000403	0.011	-53.2	-135, 38.8	0.0113	0.951
Female	-0.316	-.305	0.330	3.48	-2.99, 9.95	0.319	0.944
Patient	-0.872	-0.887	0.445	-1.64	-4.81, 1.52	0.428	0.934
Relative	-0.0250	-0.00386	0.397	84.6	-13.8, 183	0.385	0.948
Level	-1.72	-1.74	0.218	-1.29	-2.07, -0.505	0.214	0.961
Intercept	-1.56	-1.59	0.362	-2.09	-3.53, -0.645	0.350	0.950
$\sigma(u_0)$	1.36	1.33	0.289	-1.97	-3.29, -0.649	0.284	0.976

too high. We conclude that the quadrature estimates can be relied on in this case whereas the PQL estimates appear to be biased. The standard errors estimated by quadrature appear to be correct whereas those estimated by PQL may be too low, although we did not investigate this directly.

Discussion

Multilevel models are useful for investigating differences in means and variances between groups when some of the subjects are related to each other or otherwise clustered in groups. Other methods such as within cluster or between cluster estimators are less efficient because they use only part of the information and inferences based on ordinary regression are invalid. Unlike the other methods, multilevel models also provide estimates of variance components which were of interest in their own right in the present study. However, multilevel models must be used with care. Assuming a normal distribution of the random effects can lead to incorrect conclusions. This may be avoided by inspecting the empirical Bayes estimates of the residuals at each level or by using non-parametric maximum likelihood. When the response variable is non-normal, we have to make further assumptions regarding transformations, distribution, links and level-1 variances. Estimation of the model becomes approximate only so that it is necessary to assess the accuracy of the parameter estimates. When Gaussian quadrature is used, it is straightforward to assess the parameters because we can compare solutions with increasing numbers of quadrature points. So, although there are examples where quadrature falls down due to numerical instability, this is usually detectable by comparing estimates with different numbers of points. For quasilikelihood methods, we can also compare increasingly accurate approximations (first order MQL through to second order PQL) to assess convergence, but if the change in parameters is not negligible, the approximation cannot be refined any further except when higher order approximations are available. Instead we can employ parametric bootstrap estimation, a very computer intensive procedure. Another method of parameter estimation, Markov Chain Monte Carlo (MCMC), is also very computer intensive and convergence needs to be carefully checked. To date, no single approach to parameter estimation for GLMMs is available that is suitable for all situations and the development of new methods is currently an active research area (see e.g., McCulloch & Searle, 2001, and references therein).

References

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, *55*, 117-128.

Bock, R. D. & Lieberman, M. (1972). Fitting a response model for n dichotomously scored items. *Psychometrika*, *33*, 179-97.

Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*, 9-25.

Breslow, N. E. & Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika*, *82*, 81-91.

Davies, R. (1987). Mass point methods for dealing with nuisance parameters in longitudinal studies. In R. Chrouchly (Ed.), *Longitudinal data analysis* (pp. 88-109). Aldershot: Averbury.

Davies, R. & Pickles, P. (1987). A joint trip timing store-type choice model for grocery shopping, including inventory effects and nonparametric control for omitted variables. *Transportation Research*, *21A*, 345-61.

Diggle, P. J. (1988). An approach to the analysis of repeated measures. *Biometrics*, *44*, 959-71.

Diggle, P. J., Liang, K-Y. & Zeger, S. L. (1996). *Analysis of longitudinal data*. Oxford: Oxford University Press.

Faraone, S. V., Seidman, L. J., Kremen, W. S., Pepple, J. R., Lyons, M. J. & Tsuang, M. T. (1995). Neuropsychological functioning among the nonpsychotic relatives of schizophrenic patients: A diagnostic efficiency analysis. *Journal of Abnormal Psychology, 104*, 286-304.

Faraone, S. V., Seidman, L. J., Kremen, W. S., Toomey, R., Lyons , M. J. & Tsuang, M. T. (1996). Neuropsychological functioning among the elderly nonpsychotic relatives of schizophrenic patients. *Schizophrenia Research, 21*, 27-31.

Follmann, D. A. & Lambert, D. (1989). Generalizing logistic regression by nonparametric mixing. *Journal of the American Statistical Association*, *84*, 295-300.

Goldstein, H. & Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A*, *78*, 719-27.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G., & Healy, M. (1998). *A user's guide to MLwiN*. Institute of Education, London: Multilevel Models Project.

Griffiths, T. D., Sigmundsson, T., Takei, N., Frangou, S., Birkett, P. B., Sharma, T., Reveley, A. M., & Murray, R. M. (1998). Minor physical anomalies in familial sporadic schizophrenia: The Maudsley family study. *Journal of Neurology Neurosurgery and Psychiatry*, *64*, 56-60.

Heckman, J. & Singer, B. (1984). A method of minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, *52*, 271-320.

Hsiao, C. (1986). *Analysis of panel data*. Cambridge: Cambridge University Press.

Kendler, K. S., McGuire, M., Gruenberg, A. M., O'Hara, A., Spellman, M., & Walsh, D. (1993). The Roscommon family study. IV. Affective illness, anxiety disorders, and alcoholism in relatives. *Archive of General Psychiatry*, *50*, 952-960.

Kremen, W. S., Seidman, L. J., Pepple, J. R., Lyons, M. J., Tsuang, M. T., & Faraone, S. V. (1994). Neuropsychological risk indicators for schizophrenia: A review of family studies. *Schizophrenia Bulletin*, *20*,103-119.

Laird, N. M. (1978). Non-parametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, *73*, 805-11

Lin, X. & Breslow, N. E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of American Statistical Association*, *91*, 1007-1016.

Lindsay, B. G., Clogg, C. C., & Grego., J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class. *Journal of the American Statistical Association, 86,* 96-107.

McCullagh, P. & Nelder, J. A. (1989). *Generalized linear models*. London: Chapman and Hall.

McCulloch, C. E. (1999). An introduction to generalized linear mixed models. In H. Fiedl, A. Berhold and G. Kauermann (Eds.), *Proceedings of the 14th International Workshop on Statistical Modelling* (pp. 39-56): Graz, Austria.

McCulloch, C. E. & Searle, S. R. (2001). *Generalized, linear, and mixed models*. New York: Wiley.

McGuffin P., Asherson P., Owen, M., & Farmer A. (1994). The strength of the genetic effect: Is there room for an environmental influence in the aetiology of schizophrenia? *British Journal of Psychiatry*, *164*, 593-599.

Morris, C. (1983). Parametric empirical Bayes inference, theory and applications. *Journal of the American Statistical Association, 78*, 47-65.

Morris, R. G., Ahmed, S., Syed, G. M., & Toone, B. K. (1993). Neural correlates of planning ability: Frontal lobe activation during Tower of London task. *Neuropsychologica*, *31*, 1367-1378.

Rabe-Hesketh, S. & Pickles, A. (1999). Generalised linear latent and mixed models. In H. Fiedl, A. Berhold and G. Kauermann (Eds.), *Proceedings of the 14th International Workshop on Statistical Modelling* (pp. 332-339): Graz, Austria.

Rabe-Hesketh, S. & Pickles, A. (2001). *Correcting for measurement error using a non-parametric exposure distribution and non-normal errors*. Submitted for publication.

Rabe-Hesketh, S., Pickles, A. & Skrondal, A. (2001). GLAMM: A general class of multilevel models and a Stata program. *Multilevel Modelling Newsletter*, *13*, 17-23.

Rabe-Hesketh, S., Pickles, A. & Taylor, C. (2000). Generalised linear latent and mixed models. *Stata Technical Bulletin,* 53, 47-57.

Rodíguez, G. & Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A*, *158*, 73-89.

Saykin A. J., Gur, R. C., Gur, R. E., Mozley, P. D., Mozley, L. H., Resnick, S. M., Kester, D. B. & Stafiniak, P. (1991). Neuropsychological function in schizophrenia. Selective impairment in memory and learning. *Archives of General Psychiatry, 48,* 618-624.

Saykin, A. J., Shtasel, D. L., Gúr, R. E., Kester, D. B., Mozley, L. H., Stafiniak, P., & Gur, R. C. (1994). Neuropsychological deficits in neuroleptic naive patients with first episode schizophrenia. *Archives of General Psychiatry, 51*,124-131.

Sham, P. C. (1998). Statistical methods in psychiatric genetics. *Statistical Methods in Medical Research*, *7*, 279-300.

Simar, L. (1976). Maximum likelihood estimation of a compound Poisson process. *Annals of Statistics*, *4*, 1200-9.

Toulopoulou, T., Morris, R. G., Rabe-Hesketh, S., King, H., & Murray, R. M. (1999). Using neuropsychology as the phenotype to explore genetic expression in schizophrenia: a study of schizophrenic patients and their first-degree relatives. *Molecular Psychiatry, 4, Supplement 1*, S28.