Module 7: Multilevel Models for Binary Responses

Concepts

Fiona Steele[1]
Centre for Multilevel Modelling

Pre-requisites

- Modules 1-3, 5, 6

Contents

[1] With many thanks to Rebecca Pillinger, George Leckie, Kelvyn Jones and Harvey Goldstein for comments on earlier drafts.

Introduction

In Module 6 we saw how multiple regression models for continuous responses can be generalised to handle binary responses. At the end of the module (C6.8), we then considered models for grouped or clustered binary data where the response variable is a proportion and the explanatory variables are defined at the group level. The application of these models was illustrated in an analysis of the proportion of voters in each state intending to vote for George Bush, including as predictors the proportion of non-white respondents in a state and the proportion who reported regular attendance at religious services.

A particular issue in the analysis of proportions is the presence of extrabinomial variation, caused by a violation of the assumption that the binary responses on which a proportion is based are independent. It was suggested in Module 6 that one way to allow for clustering (non-independence) due to omitted group-level predictors is to fit a multilevel model with group-level random effects. We pursue this approach here, but our focus is on showing how multilevel models can be applied more generally to two-level binary response data with predictors that can be defined at both level 1 and level 2.

Some examples of research questions that can be explored through multilevel models for binary responses are:

- What is the extent of between-state variation in US voting preferences (Republican vs. Democrat)? Can between-state differences in voting be explained by differences in the ethnic or religious composition of states? Do individual-level variables such as age and gender have different effects in different states?

- Does the use of dental health services (e.g. whether a person visited a dentist in the last year) vary across areas? To what extent are any differences between areas attributable to between-area differences in the provision of subsidised services or differences in the demographic and socio-economic composition of residents?

In both of the above examples, the study populations have a two-level hierarchical structure with individuals at level 1 and areas at level 2, but structures can have more than two levels and may be non-hierarchical (see Module 4). In this module, as in Module 5 for continuous responses, we consider only models for two-level hierarchical structures.

The aim of this module is to bring together multilevel models for continuous responses (Module 5) and single-level models for binary responses (Module 6). We shall see that many of the extensions to the basic multilevel model introduced in Module 5 – for example random slopes and contextual effects – apply also to binary responses. However, there are some important new issues to consider in the interpretation and estimation of multilevel binary response models.

Introduction to the Example Dataset

We will illustrate methods for analysing binary responses using data from the 2004 National Annenberg Election Study (NAES04), a US survey designed to track the dynamics of public opinion over the 2004 presidential campaign. See http://www.annenbergpublicpolicycenter.org for further details of the NAES.

In this module (as in Module 6) we analyse data from the National Rolling Cross-Section of NAES04. The response variable for our analysis is based on voting intentions in the 2004 general election (variable cRC03), which was asked of respondents interviewed between 7 October 2003 and 27 January 2004. The question was worded as follows:

- *Thinking about the general election for president in November 2004, if that election were held today, would you vote for George W. Bush or the Democratic candidate?*

The response options were: Bush, Democrat, Other, Would not vote, or Depends. A small number of respondents reported that they did not know or refused to answer the question. Don't knows and refusals were excluded from the analysis, and the remaining categories were combined to obtain a binary variable coded 1 for Bush and 0 otherwise.

In Module 6 we analysed data from three states. We now extend the analysis sample to include all 49 states in the study, containing a total of 14,169 respondents.

We consider six *individual-level* explanatory variables:

- Annual household income, grouped into nine categories (1 = less than $10k, 2 = $10-15K, 3 = $15-25K, 4 = $25-35K, 5 = $35-50K, 6 = $50-75K, 7 = $75-100K, 8 = $100-150K, 9 = $150k or more). This variable is treated as continuous in all analyses and is centred around its sample mean of 5.23

- Sex (0 = male,1 = female)

- Age in years (mean centred)

- Type of region of residence (0 = rural, 1 = urban)

- Marital status (1 = currently married or cohabiting, 2 = widowed or divorced, 3 = not currently living with a partner and never married)

- Frequency of attendance at religious services (0 = less than weekly or never, 1 = weekly or more)

and one *state-level* explanatory variable, calculated by aggregating an individual-level variable giving the frequency of attendance at religious services:

- Proportion of respondents who attend religious services at least once a week

C7.1 Two-level Random Intercept Model for Binary Responses

C7.1.1 Generalised linear random intercept model

Consider a two-level structure where a total of *n* individuals (at level 1) are nested within *J* groups (at level 2) with n_j individuals in group *j*. Throughout this module we use 'group' as a general term for any level 2 unit, e.g. an area or a school. We denote by y_{ij} the response for individual *i* in group *j*, and by x_{ij} an individual-level explanatory variable. Recall from C5.2, equation (5.4), the random intercept model for continuous *y*:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij} \qquad (7.1)$$

where the group effects or level 2 residuals u_j and the level 1 residuals e_{ij} are assumed to be independent and to follow normal distributions with zero means:

$u_j \sim N(0, \sigma_u^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$.

We can also express the model in terms of the mean or *expected value* of y_{ij} for an individual in group *j* and with value x_{ij} on *x*:

$$E(y_{ij}|x_{ij}, u_j) = \beta_0 + \beta_1 x_{ij} + u_j. \qquad (7.2)$$

For a binary response y_{ij}, we have $E(y_{ij}|x_{ij}, u_j) = \pi_{ij} = Pr(y_{ij} = 1)$ and a *generalised linear random intercept model* for the dependency of the response probability π_{ij} on x_{ij} is written:

$$F^{-1}(\pi_{ij}) = \beta_0 + \beta_1 x_{ij} + u_j \qquad (7.3)$$

where F^{-1}("*F* inverse") is the link function, taken to be the inverse cumulative distribution function of a known distribution (see C6.3.1). In Module 6, we considered three link functions: the logit, probit and complementary log-log (clog-log) functions. Here we will focus on the logit link, with some discussion of the probit, but everything we say for the logit applies equally to the other link functions.

The key point to note about (7.3) is that, although the left hand side is a nonlinear transformation of π_{ij}, the right hand side takes the same form as that of (7.2) for continuous *y*, i.e. it is linear in terms of the parameters β_0 and β_1 and the level 2 residuals u_j. Therefore this simple random intercept model for binary *y* can be extended in the same ways that we considered in Module 5 for continuous *y*, including the addition of further explanatory variables defined at level 1 or 2, cross-level interactions, and random slopes (coefficients).

C7.1.2 Random intercept logit model

In a logit model $F^{-1}(\pi_{ij})$ is the log-odds that *y* = 1 (see C6.3.2), so (7.3) becomes

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \beta_1 x_{ij} + u_j \quad (7.4)$$

where $u_j \sim N(0, \sigma_u^2)$.

Interpretation of β_0 and β_1

β_0 is interpreted as the *log-odds* that *y* = 1 when *x* = 0 and *u* = 0 and is referred to as the *overall intercept* in the linear relationship between the log-odds and *x*. If we take the exponential of β_0, $\exp(\beta_0)$, we obtain the *odds* that y = 1 for *x* = 0 and *u* = 0.

As in the single-level model, β_1 is the effect of a 1-unit change in *x* on the log-odds that *y* = 1, but it is now the effect of *x* after adjusting for (or holding constant) the group effect *u*. If we are holding *u* constant, then we are looking at the effect of *x* for individuals within the same group so β_1 is usually referred to as a *cluster-specific effect*. In C7.3 we will compare this cluster-specific effect with the effect of *x* averaging across groups (the *population-average effect*). These effects are equal for a multilevel continuous response model, so that in Module 5 we made no distinction between them, but they will not be equal for a generalised linear multilevel model (unless $\sigma_u^2 = 0$).

As in a single-level logit model, $\exp(\beta_1)$ can be interpreted as an odds ratio, comparing the odds that *y* = 1 for two individuals (in the same group) with *x*-values spaced 1 unit apart.

Interpretation of u_j

While β_0 is the overall intercept in the linear relationship between the log-odds and *x*, the intercept for a given group *j* is $\beta_0 + u_j$ which will be higher or lower than the overall intercept depending on whether u_j is greater or less than zero. As in the continuous response case, we refer to u_j as the group (random) effect, group residual, or level 2 residual. The variance of the intercepts across groups is $\mathrm{var}(u_j) = \sigma_u^2$, which is referred to as the between-group variance adjusted for *x*, the between-group residual variance, or simply the level 2 residual variance. (Quite often 'residual' is omitted and we say 'level 2 variance', but remember that if the model contains explanatory variables then σ_u^2 is always the *unexplained* level 2 variance.)

We can obtain estimates of u_j that can be plotted with confidence intervals to see which groups are significantly below or above the average of zero (a caterpillar plot). These estimates are interpreted in the same way as for continuous response models (see C5.1.2 and C5.2.2); the only difference is that in a logit model they represent group effects on the log-odds scale.

In analysing multilevel data, we are often interested in the amount of variation that can be attributed to the different levels in the data structure and the extent to which variation at a given level can be explained by explanatory variables. In Module 5 (C5.1.1) we met the *variance partition coefficient* which measures the proportion of the total variance that is due to differences between groups. There is no unique

way of defining a VPC for binary data, but we shall consider one approach in C7.2.4. (The problem is analogous to the difficulty in defining R^2 for binary data - see C6.4.)

Predicted response probabilities

As in the single-level case, we can re-organise (7.4) to obtain an expression for the response probability:

$$\pi_{ij} = \frac{\exp(\beta_0+\beta_1 x_{ij}+u_j)}{1+\exp(\beta_0+\beta_1 x_{ij}+u_j)}. \qquad (7.5)$$

(See equation (6.10) in C6.3.2 for the single-level version, i.e. without group effects.)

We can calculate the predicted response probability for individual *i* in group *j* by substituting the estimates of β_0, β_1 and u_j obtained from the fitted model as follows:

$$\hat{\pi}_{ij} = \frac{\exp(\hat{\beta}_0+\hat{\beta}_1 x_{ij}+\hat{u}_j)}{1+\exp(\hat{\beta}_0+\hat{\beta}_1 x_{ij}+\hat{u}_j)}.$$

We can also make predictions for 'ideal' or 'typical' individuals with specific combinations of *x*-values, but we also need to make a decision on what value to substitute for u_j. We will discuss predicted probabilities in C7.4.

C7.1.3 Example: Between-state variation in voting intentions in the US

We illustrate the application and interpretation of the random intercept logit model (7.4) in an analysis of voting intentions in the 2004 US general election. A two-level model is used to allow for correlation between voting intentions of individuals living in the same state, and to explore the extent of between-state variation in voting intentions.

Null model (without explanatory variables)

Table 7.1 shows the results from fitting a multilevel logit model for the probability of voting for Bush with state random effects but no explanatory variables. This 'null' model is sometimes referred to as a variance components model. The odds of voting Bush for an 'average' state (with u_j = 0) are estimated as exp(-0.107) = 0.90, and the corresponding probability is 0.9/(1+0.9) = 0.47.

Table 7.1. Multilevel logit model for voting Bush, with state effects, US 2004

Parameter	Estimate	Standard error
β_0 (Constant)	-0.107	0.049
σ_u^2 (Between-state variance)	0.091	0.023

The between-state variance in the log-odds of voting Bush is estimated as 0.091 with a standard error of 0.023. There are various ways that we might test the significance

of the between-state variance, and the approaches available to us depend on the algorithm used to fit the model. We discuss algorithms and software in C7.7 and in the Technical Appendix. Ideally we would use a likelihood ratio test (as in the continuous response case), but this option is only available when maximum likelihood estimation is used. Because the estimates in Table 7.1 were obtained using a quasi-likelihood procedure[2], we will use a Wald test. The Wald test was described in C6.5.5 for testing coefficients in a single-level model, but it can be used to test hypotheses about any model parameter. When used to test a hypothesis about a variance parameter (e.g. the between-state variance), the test is crude because it depends on the questionable assumption that the variance estimate is normally distributed.[3]Nevertheless, it will give us some indication of the strength of the evidence for state effects. The Wald test statistic is the square of the Z-ratio, i.e. $(0.091/0.023)^2 = 15.65$ which is compared with a chi-squared distribution on 1 degree of freedom, giving a p-value less than 0.001. We therefore conclude that there is significant variation between states in the proportion who intend to vote for Bush.

Another issue to consider when testing variance parameters is that variances are by definition non-negative. The null hypothesis is that $\sigma_u^2 = 0$, but the alternative hypothesis is one-sided ($\sigma_u^2 > 0$) rather than two-sided ($\sigma_u^2 \neq 0$). One suggested approach to the problem is to halve the p-value obtained from comparing the likelihood ratio statistic with a chi-squared distribution (see Snijders and Bosker (1999, Section 6.2) for a discussion). Note that the above applies to tests of variance parameters in multilevel models for any type of outcome variable, not just binary *y*.

σ_u^2 is the between-state variance in the log-odds of voting Bush, but it is difficult to assess the size of the state effects when using the log-odds scale. Instead we can calculate predicted probabilities of voting Bush, using (7.5) with no *x*-variable, assuming different values for the state effect u_j. We have already calculated the predicted probability for an 'average' state with $u_j = 0$. Under the assumption that u_j follow a normal distribution, we would expect approximately 95% of states to have a value of u_j within 2 standard deviations of the mean of zero, i.e. between approximately $-2\hat{\sigma}_u = -2\sqrt{0.091} = -0.603$ and +0.603. This type of interval is sometimes called a *coverage interval*. Substituting in (7.5) these values for u_j and our estimate for β_0 from Table 7.1 we obtain the following predictions.

For a state 2 standard deviations below the mean:

$$\hat{\pi} = \frac{\exp(-0.107-0.603)}{1+\exp(-0.107-0.603)} = 0.33$$

[2] Second order penalized quasi-likelihood (PQL2) – see C7.7 and the Technical Appendix for details.
[3]We are referring here to the sampling distribution of the estimated variance. Imagine taking repeated samples of respondents within states, and fitting a multilevel logit model to each sample. You will get a different estimate of the between-state variance each time. The distribution of this variance estimate across samples is the sampling distribution which, in a Wald test, is assumed normal. The sampling distribution of a variance estimate is in fact positively skewed (the right tail of the distribution is longer) because variances must be greater than zero. The Wald test performs particularly poorly when the level 2 variance estimate is close to its boundary of zero.

For a state 2 standard deviations above the mean:

$$\hat{\pi} = \frac{\exp(-0.107 + 0.603)}{1 + \exp(-0.107 + 0.603)} = 0.62$$

We would therefore expect the proportion voting Bush to lie between 0.33 and 0.62 in the middle 95% of states.

We now examine estimates of the state effects, $\hat{u}_j$, obtained from the null model. Figure 7.1 is a 'caterpillar plot' with the state effects shown in rank order together with 95% confidence intervals. This plot is interpreted in the same way as for a continuous response model (see C5.1.2), but the level 2 residuals are now state effects on the log-odds scale. As before, a state whose confidence interval does not overlap the line at zero (representing the mean log-odds of voting Bush across all states) is said to differ significantly from the average at the 5% level. In this case, many of the confidence intervals include zero and there are no obvious outliers with especially large $\hat{u}_j$. The three states with the lowest probability of voting Bush (largest negative values of $\hat{u}_j$) are Washington DC, Rhode Island and Massachusetts, while the three with the highest response probability (largest positive values of $\hat{u}_j$) are Utah, Montana and Texas. Note that a few states have very narrow intervals; these are the states with the largest samples sizes, including California, New York and Texas which were studied in Module 6.

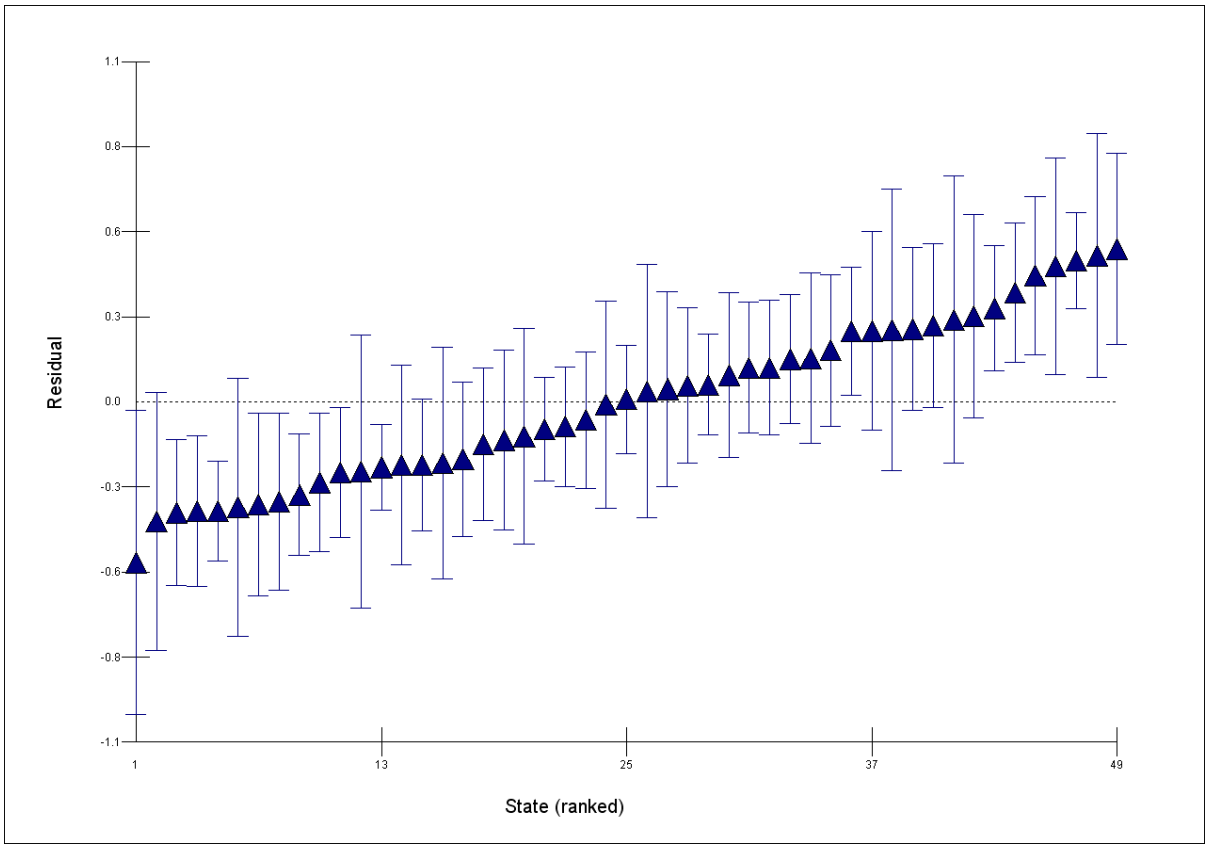

Figure 7.1. Caterpillar plot showing state residuals with 95% confidence intervals for log-odds of voting Bush

Fitting the null model to proportions

The results in Table 7.1 were obtained by fitting a multilevel model to individual-level binary voting intentions. Given that the model does not include any individual-level explanatory variables, however, we would have obtained exactly the same results by analysing the proportions of Bush voters in each state – a dataset with 49 observations rather than 14,169. Multilevel modelling of grouped binary data, with a random effect for each group, was discussed in C6.8.4 as a way of allowing for omitted group-level predictors which might lead to extrabinomial variation. In general, however, the groups for which proportions are available need not correspond to level 2 units in a multilevel model. For example, it is common to have mortality rates for areas that are broken down by age and sex. In that case, the 'group' would correspond to a particular age-sex combination within an area and the areas would form the level 2 units; we therefore fit a two-level model with age-sex categories at level 1 and areas at level 2. Conceptually, we have a two-level structure with individuals within areas but, because we do not have information to distinguish individuals within a given age-sex category, there is nothing to be gained from expanding the data to give an observation for each individual; it is more efficient to store and analyse the data in aggregate form.

Adding an explanatory variable: a random intercept model

We next consider adding an explanatory variable, household annual income, which ranges from 1 to 9 and has been centred around its sample mean (across all individuals regardless of their state) of 5.23. The results from fitting a random intercept model are given in Table 7.2. The intercept estimate of -0.099 is now the estimated log-odds of voting Bush for an individual with a mean household income living in an 'average' state. There is a highly significant, positive income effect (Z= 0.140/0.008 = 17.5). Controlling for state differences, we would expect the odds of voting Bush to increase by a factor of exp(0.140) = 1.15 for each 1-unit increase in income, i.e. a 15% increase. We would therefore expect the odds of voting Bush to be exp(8×0.14) = 3.1 times higher for an individual in the highest income band (coded 9) than for an individual in the same state but in the lowest income band (coded 1).[4]

Table 7.2. Multilevel logit model for voting Bush, with state and income effects, US 2004

Parameter	Estimate	Standard error
β_0 (Constant)	-0.099	0.056
β_1 (Income, centred)	0.140	0.008
σ_u^2(Between-state variance)	0.125	0.030

Figure 7.2 shows the predicted state lines for the relationship between the log-odds of voting Bush and household annual income. Note that some lines are shorter than others; this is because in some states there are no respondents in the highest income band. The intercept of the line for state *j* is estimated as -0.099 + $\hat{u}_j$. As expected for a random intercept model, the lines are parallel because we have assumed that

[4] The odds ratio comparing an individual in band 9 with an individual in band 1 could also be calculated as exp(8×0.14) = $[\exp(0.14)]^8$ = 1.15^8= 3.1

the effect of income is the same for each state. We will relax this assumption in C7.5 by introducing a random coefficient for income.

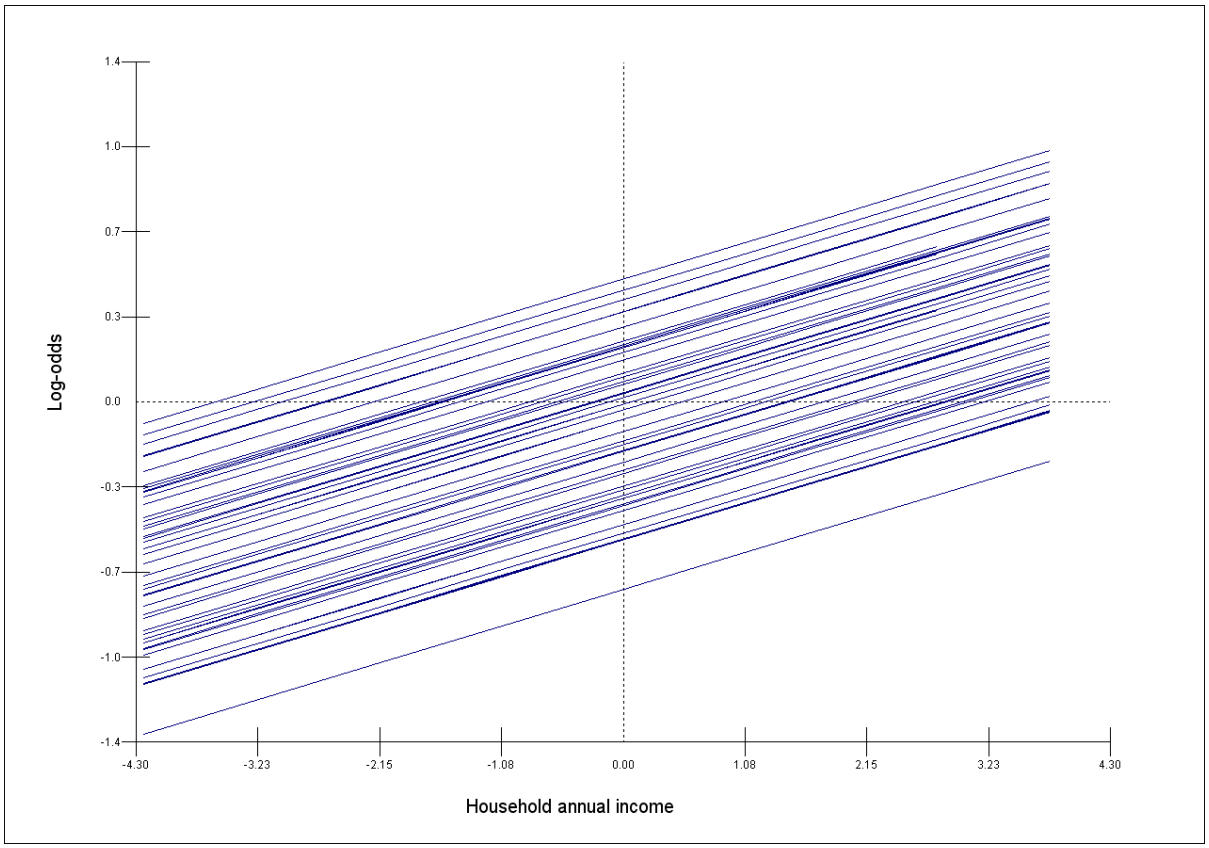

Figure 7.2. Predicted state lines for the relationship between the log-odds of voting Bush and household annual income, US 2004

C7.2 Latent Variable Representation of a Random Intercept Model for Binary Responses

In C6.4 we saw how a generalised linear model for a single-level binary response *y* can be re-expressed in terms of a linear model for a continuous latent (unobserved) variable *y**, where *y** represents the underlying propensity of being in response category 1 rather than 0. In this section we describe a threshold model for a two-level structure which is an alternative representation of the generalised linear random intercept model given by (7.3). There are several advantages of viewing the model in this way, all of which will be discussed in this section:

i) The threshold model allows us to see more easily the relationship between the logit and probit models

ii) We can see more clearly the impact of adding random effects on the coefficients of a multilevel model (the *scaling* issue)

iii) We can see the impact of adding level 1 explanatory variables on the estimate of the level 2 variance σ^2_u

iv) We can obtain directly a variance partition coefficient that measures the proportion of variance in *y** due to group (level 2) effects

C7.2.1 Two-level random intercept threshold model

Suppose that underlying the binary response y_{ij}there is a continuous latent variable y^*_{ij} that is related to the observed y_{ij} as follows:

$$y_{ij} = \begin{cases} 1 & \text{if } y^*_{ij} \geq 0 \\ 0 & \text{if } y^*_{ij} < 0 \end{cases}$$

As in the single-level case (C6.4) the threshold or cutpoint is arbitrary because y^*_{ij} is unobserved; we have chosen a zero threshold.

We can define a two-level random intercept model for y^*_{ij}, just as for any continuous variable:

$$y^*_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e^*_{ij} \qquad (7.6)$$

where e^*_{ij} is a level 1 residual with mean zero and variance σ^2_{e*}. Because y^*_{ij} is unobserved, we need to set its scale which we do by fixing σ^2_{e*}.We also need to specify the distribution of e^*_{ij}. Assuming a normal distribution with σ^2_{e*} = 1 leads to a random intercept probit model, while a logistic distribution with σ^2_{e*} = 3.29 leads to a random intercept logit model. The logit form of the model is sometimes called a *logistic-normal* model because the level 1 residual is assumed to follow a logistic distribution, while the level 2 residual is assumed normal.

As in the single-level case (C6.4), the relationship between the coefficients from random intercept logit and probit models is approximately

$$\hat{\beta}_{\text{logit}} \approx \sqrt{3.29}\ \hat{\beta}_{\text{probit}} = 1.8\ \hat{\beta}_{\text{probit}}$$

although empirical research suggests that a scaling factor between 1.6 and 1.7 gives a better approximation.

C7.2.2 Comparison of a single-level and multilevel threshold model

Recall the single-level version of the threshold model (C6.4):

$$y_i^* = \beta_0 + \beta_1 x_i + e_i^*. \qquad (7.7)$$

Because $\sigma_u^2 \geq 0$ and the variance of e_{ij}^* is fixed – at 3.29 or 1 according to whether a logit or probit link is used – the residual variance in y_{ij}^* for the multilevel model (7.6) will always be greater than or equal to the residual variance in its single-level counterpart (7.7). Specifically $\text{var}(y_{ij}^*) = \sigma_u^2 + \sigma_{e*}^2$ for a multilevel model and $\text{var}(y_i^*) = \sigma_{e*}^2$ for a single-level model. The increase in the residual variance when a random effect is added to the model stretches the scale of *y** which means that the coefficients β_0 and $\beta_1$will be scaled up. Therefore coefficients from a random intercept model will be greater in magnitude than coefficients from its single-level version, provided that the distribution of each explanatory variable is the same across groups (we return to this proviso in a moment). For a logit model, for example, the relationship between the random intercept (RI) and single-level (SL) estimates is approximately:

$$\hat{\beta}_1^{\text{RI}} \approx \hat{\beta}_1^{\text{SL}} \sqrt{\frac{\sigma_u^2 + 3.29}{3.29}}. \qquad (7.8)$$

Replacing 3.29 by 1 gives the corresponding relationship between the random intercept and single-level coefficients from probit models.

It is important to note that this scaling issue does not arise in continuous response models because the level 1 residual variance is not fixed: the total residual variance in a multilevel model for continuous *y* will be approximately equal to the residual variance in the corresponding single-level model. Note also that adding random effects to a binary response model does not increase the amount of variability in the *data*, i.e. in *y*. It is the residual variance in the unobserved *y**that is increased which is allowed because, as a latent variable, its scale is arbitrary. The scaling we have chosen fixes the part of the residual variance in *y** that is allocated to level 1, while the level 2 variance is a free parameter which we estimate.

To demonstrate the impact of adding a random effect on the estimated coefficients of a logit model, single-level and random intercept models were fitted to simulated data. The data have a two-level structure with 1000 groups at level 2 and 30 level 1 units within each group, giving a total sample size of 30,000. Two level 1 explanatory variables were generated: x_1 (a normally distributed variable with mean zero and

variance one) and x_2 (a binary variable with approximately 50% in each category). We then generated a binary response that depends on x_1 and x_2 via a random intercept logit model with a level 2 variance of 1. An important feature of the simulation is that x_1 and x_2 have the same distribution within each level 2 unit, i.e. the mean and standard deviation of x_1 is approximately the same for each of the 1000 level 2 units and the proportion in each category of x_2 is approximately 50% in each level 2 unit; this implies that x_1 and x_2 are both uncorrelated with the level 2 random effect.

Table 7.3 shows the estimated coefficients and standard errors from fitting single-level and random intercept logit models to the simulated data, and the final column gives the ratio of the random intercept (RI) coefficient to the single-level (SL) coefficient. Substituting $\widehat{\sigma}_u^2 = 1.018$ into (7.8) gives an expected RI:SL ratio of 1.14. The actual ratio is somewhat larger than this, but is similar for both β_1 and β_2. As in the comparison of logit and probit coefficients (see C6.4) the relationship given by (7.8) is only an approximation.

Table 7.3. Estimates from single and random intercept logit models, simulated data

	Single-level (SL)		***Random intercept (RI)***		
Parameter	*Est.*	*SE*	*Est.*	*SE*	*RI/SL*
β_0 (Constant)	0.221	0.017	0.257	0.037	1.163
β_1 (x_1)	0.430	0.013	0.519	0.014	1.207
β_2 (x_2)	0.498	0.024	0.613	0.027	1.231
σ_u^2 (Level 2 variance)	-	-	1.018	0.054	-

We now compare single-level and random intercept model estimates for a real dataset. Table 7.4 gives estimates from models fitted to an educational dataset where the binary response is an indicator of whether a student achieved a score that was higher than the overall mean in an examination taken at age 16. (The binary variable was constructed from a continuous score for the purposes of this illustration.) The data have a two-level structure with 4059 students nested within 65 schools. There are four explanatory variables: a prior attainment score, a dummy for sex (coded 1 for a girl and 0 for a boy) and two dummy variables for a categorical indicator of verbal reasoning (taking 'low' as the reference). From (7.8) we would expect a RI:SL ratio of 1.070, which is close to the ratios actually estimated for all coefficients apart from β_2(Girl). However, a closer examination of the data reveals that there is substantial variation in the proportion of girls across schools, largely because the sample contains single-sex schools. Restricting the analysis to mixed-sex schools leads to the results in Table 7.5. We now expect the RI:SL ratio to be 1.083. Again the ratio for Girl is rather different from this, but at least the random intercept coefficient is now larger than the single-level coefficient as expected.

Table 7.4. Estimates from single-level and random intercept models for probability of obtaining score greater than the mean, all schools

	Single-level (SL)		***Random intercept (RI)***		
Parameter	*Est.*	*SE*	*Est.*	*SE*	*RI/SL*
β_0 (Constant)	0.708	0.098	0.828	0.188	1.169
β_1 (Prior attainment)	0.909	0.057	0.951	0.061	1.046
β_2 (Girl)	0.221	0.075	0.193	0.103	0.873
β_3 (Verbal reasoning mid)	-0.956	0.101	-1.030	0.105	1.077
β_4 (Verbal reasoning high)	-1.717	0.181	-1.805	0.188	1.051
σ^2_u (School-level variance)	-	-	0.499	0.107	-

Table 7.5. Estimates from single-level and random intercept models for probability of obtaining score greater than the mean, mixed schools only

	Single-level (SL)		***Random intercept (RI)***		
Parameter	*Est.*	*SE*	*Est.*	*SE*	*RI/SL*
β_0 (Constant)	0.406	0.124	0.457	0.185	1.126
β_1 (Prior attainment)	1.050	0.080	1.094	0.084	1.042
β_2 (Girl)	0.121	0.101	0.174	0.111	1.438
β_3 (Verbal reasoning mid)	-0.803	0.130	-0.865	0.137	1.077
β_4 (Verbal reasoning high)	-1.311	0.241	-1.380	0.252	1.053
σ^2_u (School-level variance)	-	-	0.572	0.164	-

We have seen that part of the difference between the single-level and random intercept coefficients is down to scaling, but the two sets of coefficients also have a different interpretation. We will return to this issue in the next section but, briefly, the coefficients from a random intercept model have a *cluster-specific* (or unit-specific) interpretation. For example, the coefficient of Girl in Table 7.5 represents a comparison between girls and boys *in the same school* (and with the same prior attainment and verbal reasoning). In the single-level model, the coefficient of Girl compares girls and boys before taking account of school effects, so it represents a comparison that is averaged across all schools (again, controlling for prior attainment and verbal reasoning). Coefficients from single-level models are said to have a *population averaged* (or *marginal*) interpretation.

C7.2.3 Impact of adding a level 1 explanatory variable to a two-level model

In a continuous response model, the addition of a level 1 explanatory variable *x* will lead to reductions in the level 1 residual variance and in the total residual variance. The coefficients of other variables in the model will also change unless the added variable is uncorrelated with each of the other variables. Most often, the

coefficients of variables with which the new variable x is correlated will decrease in magnitude.

From the threshold representation of a binary response model, however, we see that the level 1 residual variance σ^2_{e*} is fixed and can therefore not decrease; only the level 2 residual variance σ^2_u can change. Rather than decrease σ^2_{e*}, the addition of *x* will tend to increase the proportion of the total residual variance that is at level 2 (the VPC) or, equivalently, increase the ratio of the level 2 residual variance to the level 1 residual variance. This is achieved by increasing σ^2_u which, in turn, stretches the scale of the latent response *y**, leading to an increase in the absolute value of the coefficients of the explanatory variables already in the model. (The reasons for this are the same as those given in C7.2.2 above for the increase in the coefficients when random effects are added to a single-level model.) To illustrate this point, we return to the simulated data analysed in C7.2.2 and consider three models: Model 1 with x_1 only, Model 2 with x_2 only and Model 3 including both x_1 and x_2 (see Table 7.6). We find that adding x_2 to Model 1, or x_1 to Model 2, leads to an increase in the estimates of both the level 2 variance and the coefficient of the other explanatory variable. The coefficient and the level 2 standard deviation increase by a similar factor. For example, the ratio of the estimate of β_2 in Model 3 to its estimate in Model 2 is 0.613/0.573 = 1.070, while the ratio of the estimates of σ_u is $\sqrt{1.018/0.914} = 1.055$.

Table 7.6. Impact of adding predictors to a random intercept logit model, simulated data

	Model 1		***Model 2***		***Model 3***	
Parameter	*Est.*	*SE*	*Est.*	*SE*	*Est.*	*SE*
β_0 (Constant)	0.550	0.034	0.254	0.035	0.257	0.037
β_1 (x_1)	0.507	0.014	-	-	0.519	0.014
β_2 (x_2)	-	-	0.573	0.026	0.613	0.027
σ^2_u (Level 2 variance)	0.976	0.052	0.914	0.049	1.018	0.054

In practice, with real data, the coefficients and the level 2 standard deviation will not all increase by similar factors when a new variable is added to the model. Furthermore, both coefficients and the level 2 variance may decrease. The simple pattern demonstrated above arises only in the unusual situation where the added variable is uncorrelated with the other explanatory variables and is evenly distributed across level 2 units. Remember that the data analysed in Table 7.6 were generated in such a way thatx_1 and $x_2$are uncorrelated with each other and their distributions are approximately the same across level 2 units (see C7.2.2). In analysing real data, coefficients and level 2 variances may increase or decrease in absolute value as variables are added to the model. The key point to bear in mind is that such changes need to be interpreted with caution because they will, at least in part, be an artefact of the scaling of *y**.

Further discussion of the issues raised in this section and C7.2.2 can be found in Snijders and Bosker (1999, Section 14.3.5).

C7.2.4 Variance partition coefficient in terms of *y**

Recall from C5.1.1 the following formula for the variance partition coefficient:

$$\text{VPC} = \frac{\text{level 2 residual variance}}{\text{level 1 residual variance + level 2 residual variance}} \qquad (7.9)$$

which measures the proportion of the total residual variance that is due to between-group variation.

As we saw in C7.2.2, we can express a generalised linear random intercept model as a threshold model for the latent propensity y_{ij}^* to be in response category 1. This model has the same form as a random intercept model for a continuous response with the level 2 and level 1 residuals on the same scale, that is both influencing y_{ij}^*, so (7.9) becomes:

$$\text{VPC} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_{e*}^2} \qquad (7.10)$$

where σ_{e*}^2 = 1 for a probit model and 3.29 for a logit model.

The VPC given by (7.10) is interpreted as the proportion of the total residual variance in the propensity to be in response category 1 that is due to differences between groups.

Applying (7.10) to the random intercept model of voting Bush (see Table 7.2), we obtain VPC = 0.125/(0.125 + 3.29) = 0.037. Adjusting for the effects of income, almost 4% of the remaining variance in the propensity to vote Bush is due to between-state variation.

C7.3 Population-Averaged and Cluster-Specific Effects[5]

In the previous sections, we have seen how a generalised linear model for binary data can be extended to allow for dependencies that result when individuals are nested within groups (or clusters). The multilevel random intercept model introduced in C5.2 for a continuous response *y* can be extended to handle binary responses, where group-specific random effects allow for unobserved group characteristics that lead to within-group dependencies. However, the random effects approach is just one way of allowing for clustering. In this section, we describe an alternative approach – marginal or population-averaged models –and discuss the interpretation of coefficients estimated from these and random effects models. Further discussion and references can be found in Hedeker and Gibbons (2006, Chapter 8).

C7.3.1 Marginal model for clustered binary data

A marginal model consists of two components: (i) a generalised linear model specifying the relationship between the response probability π_{ij} and predictors x_{ij}, and (ii) a specification of the structure of the correlations between pairs of individuals in the same group. As usual, we assume that the binary responses y_{ij} follow a Bernoulli distribution with mean π_{ij} and variance $\pi_{ij}(1 - \pi_{ij})$.

The first component of a marginal model is simply the generalised linear model described in Module 6 for single-level data, but we add *i* and *j* subscripts toπ and *x*to reflect the two-level structure of the data:

$$\mathrm{F}^{-1}(\pi_{ij}) = \beta_0 + \beta_1 x_{ij} \qquad (7.11)$$

A number of structures can be specified for the within-group correlations, but the most popular choices are:

- *Independence* – zero pairwise correlations between individuals in the same group, which is equivalent to fitting a single-level model
- *Exchangeable*(also known as *compound symmetry*) – non-zero, but equal correlations between each pair of individuals in the same group; equivalent to the within-group correlation structure assumed in a random intercept model
- *Autocorrelation* - used in longitudinal designs where the correlation between a pair of observations (level 1 units) on an individual (level 2) depends on the length of time between them
- *Unstructured* – where all the pairwise correlations are estimated, rather than assuming a particular structure; appropriate in situations where the units

[5] Thanks to Kelvyn Jones for contributing material to this section.

within a group are ordered in some way (e.g. as in longitudinal designs)but practically feasible only when the number of observations per group is small

Note that for all of the above structures, the within-group correlation structure is assumed to be the same for each group.

A particular advantage of the marginal model is that estimates of the coefficients in (7.11) and their standard errors are fairly robust to misspecification of the correlation structure, so the choice is not crucial.

Marginal models are usually estimated using a method called Generalised Estimating Equations (GEE), and the models themselves are sometimes called GEE models. They are most commonly used in longitudinal data analysis where the clustering arises from having multiple observations over time on the same individual.

Comparison of the marginal and multilevel approaches

One major difference between the marginal and multilevel approaches to analysing clustered data is that a marginal model treats the clustering as a feature that needs to be taken into account, but which is not something of intrinsic interest. In a marginal model, the dependency between members of the same cluster is treated as a nuisance and there is no parameter representing the between-cluster variance that we can estimate, nor can we obtain estimates of cluster effects. In contrast, the between-cluster variance is an important component of a multilevel model and there is potential to allow this variance to depend on explanatory variables (via random coefficients – see C5.3 and C7.5) and to allow for further levels of clustering and non-hierarchical structures (see Module 4).

Comparing equations (7.11) and (7.3) for the marginal and random intercept models, the most obvious difference between them is that the random intercept model includes a group-level random effect. It is this random effect that allows for unobserved group-level variables and, therefore, dependency among individuals in the same group. These random effects are built into the linear predictor, while the marginal approach involves specifying a separate model component to allow for within-group dependencies. This difference between (7.11) and (7.3) means that the coefficients from the two models have a different interpretation.

C7.3.2 Interpretation of coefficients from random effects and marginal models

In the following discussion of the interpretation of marginal and random intercept models we will consider the logit form of each model, but the differences in interpretation apply equally to probit and clog-log models. To distinguish between the coefficients of the two models, we use the superscripts PA (for *population-averaged*) in the marginal model and CS (for *cluster-specific*) in the random effects model.

It should be noted that although the coefficients from marginal (population-averaged) and random effects (cluster-specific) models have different

interpretations, it is possible to calculate predicted probabilities from a random effects model that have a population-averaged interpretation (see C7.4).

The random intercept logit model with a single explanatory variable x_{ij} can be written

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0^{CS} + \beta_1^{CS}x_{ij} + u_j \qquad (7.12)$$

while the corresponding marginal logit model is

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0^{PA} + \beta_1^{PA}x_{ij}. \qquad (7.13)$$

Cluster-specific effects

β_1^{CS} is interpreted as the effect of a 1-unit change in *x* on the log-odds that *y* = 1 *for a given cluster*, that is holding constant (or conditioning on) cluster-specific unobservables captured by the random effect u_j. In other words, β_1^{CS} represents a contrast between two individuals in the same cluster (or in different clusters with the same random effect value) with *x*-values one unit apart. In C7.1.3, for example, the coefficient of household income (the only variable in the random intercept model) compares two individuals living in the same state but in adjacent income bands.

In longitudinal designs, where the clusters are individuals, the coefficient β_1^{CS} is often referred to as the *subject-specific* or *unit-specific* effect. The subject-specific effect of *x* is the effect of a 1-unit change in *x* for a given individual, that is holding constant the combination of unobserved individual characteristics represented by the random effect. This interpretation is particularly useful for time-varying (level 1) predictors. Suppose, for example, that a longitudinal study is designed to assess cancer patients' tolerance to different doses of chemotherapy.[6] Our binary response y_{ij} indicates whether patient *j* has an adverse reaction at time *i* to chemotherapy dose x_{ij} (which can vary over time for a given individual), and we analyse the data using the random intercept model (7.12). The coefficient β_1^{CS} represents the effect on a patient's risk of developing an adverse reaction of increasing the chemotherapy dose by one unit. Specifically, β_1^{CS} is the effect of *x* holding constant the time-invariant individual characteristics represented by u_j. Under certain conditions (e.g. if patients were assigned to the different doses at random rather than according to their expected reaction), β_1^{CS} may be interpreted as a *causal* effect. The random effect allows for individual differences in the tolerance to chemotherapy, and we would expect quite a large between-individual variance σ_u^2.

We might be able to explain some of this between-individual variation by including individual-level (level 2) variables in the model. Suppose, for example, that previous

[6]This example is taken from lecture notes by Charles E. McCulloch (Division of Biostatistics, University of California San Francisco) which can be downloaded from
http://psg-mac43.ucsf.edu/ticr/syllabus/courses/8/2004/05/11/Lecture/notes/lec18%20repeated%20measures%204.pdf

research has found that men and women have different tolerances which leads us to include a gender dummy x_{2j} with coefficient β_2^{CS}. Because gender is fixed over time, it does not make sense to talk about a within-person effect of gender. Instead, β_2^{CS} compares men and women with the same value for x_{1ij} and the same random effect value, i.e. the same combination of unobserved time-invariant characteristics. For level 2 variables that could potentially be manipulated (e.g. assignment to different types of drug), the cluster-specific effect may be of interest because it represents a comparison of two individuals who differ only on one explanatory variable (e.g. type of drug) but who are the same in all other respects. For a variable such as gender, however, we are more often interested in a comparison of men and women averaging across unobserved characteristics in the population, in which case we can derive population-averaged predicted probabilities by gender (see C7.4 below).

Population-averaged effects

Turning to the marginal model, β_1^{PA} is interpreted as the effect of a 1-unit change in *x* on the log-odds that *y* = 1*in the study population*, rather than within a cluster. In a marginal model we are 'averaging' across the cluster-level random effects. For example, in the above chemotherapy example with dose as the only predictor, β_1^{PA} simply represents a comparison of individuals whose dosage differs by one unit. We are 'averaging' over the between-individual differences in drug tolerance that are due to time-constant unobserved characteristics. Similarly, if gender is added as a second predictor, β_2^{PA} represents the gender effect on the risk of an adverse reaction for individuals given a specific dose of chemotherapy. Such comparisons may be useful for public health purposes where we are interested in the overall effect for the population sampled.

Comparison of cluster-specific and population-averaged coefficients

Figure 7.3 shows predicted probabilities from fitting a marginal logit model with exchangeable correlation structure and a random intercept logit model to simulated two-level data with a single explanatory variable *x*. The population-averaged predictions from the marginal model are given by the solid curve in the middle. Cluster-specific predictions were made for three values of the random effect: $2\hat{\sigma}_u$ (top curve), 0 (middle) and $-2\hat{\sigma}_u$ (bottom). We would expect the cluster-specific prediction curves to lie between the top and bottom curves of Figure 7.3 for roughly 95% of clusters.

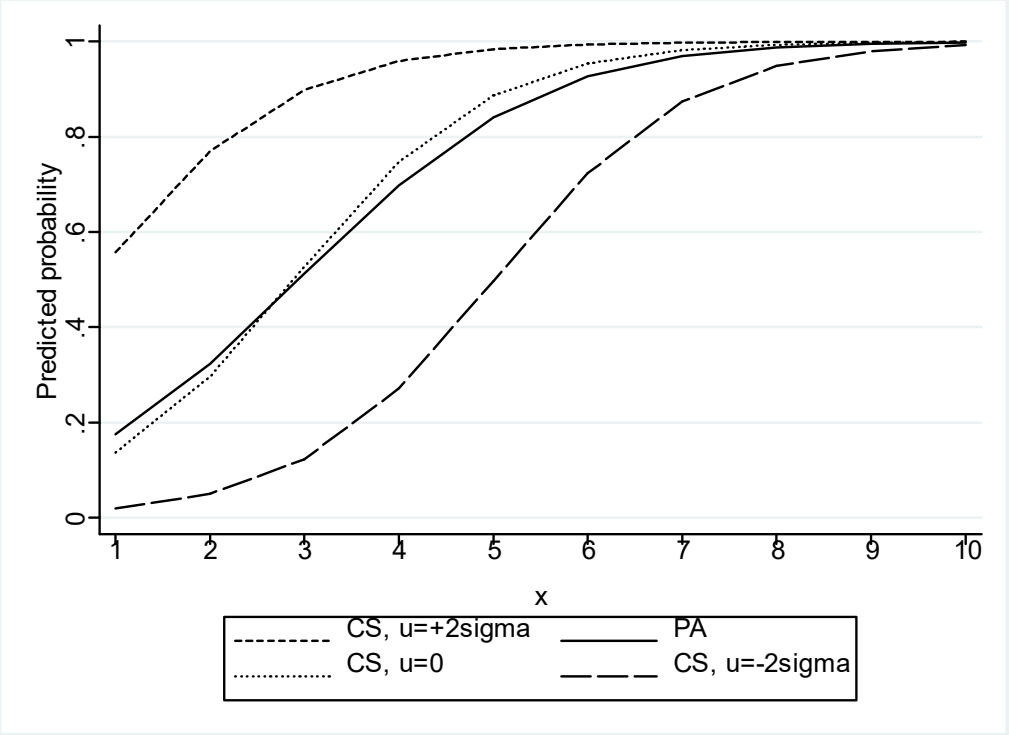

Figure 7.3. Predicted probabilities from cluster-specific and population-averaged coefficients

The cluster-specific curve at $u_j = 0$ (showing the median prediction for a given x, see C7.4) and the population-averaged curve (showing mean predictions) will always intersect at a predicted probability of 0.5. For predicted probabilities less than 0.5 the median curve is lower than the population-averaged curve, while for probabilities greater than 0.5 the median curve is the higher. Thus, for a given range of x, the range in predicted probabilities will be greater from the cluster-specific model than for the population-averaged model, implying that $\left|\hat{\beta}_1^{\mathrm{CS}}\right| > \left|\hat{\beta}_1^{\mathrm{PA}}\right|$. The curves will become further apart as the between-cluster variance $\hat{\sigma}_u^2$ in the random intercept model increases. More formally, it can be shown that the relationship between the CS and PA coefficients for any explanatory variable in a logit model is approximately:

$$\beta^{\mathrm{CS}} \approx \sqrt{\frac{\sigma_u^2+3.29}{3.29}}\beta^{\mathrm{PA}}. \quad (7.14)$$

To obtain the relationship between coefficients from CS and PA probit models, simply replace 3.29 in (7.14) by 1.

Note that when there is no clustering, $\sigma_u^2 = 0$ and $\beta^{\mathrm{PA}} = \beta^{\mathrm{CS}}$.

Given their different interpretations, we might ask the question: which of β_1^{CS} and β_1^{PA} is the more useful? The cluster-specific effect, β_1^{CS}, may be more useful in a clinical setting where there is interest in how an individual's risk of an adverse reaction would be expected to change with dose. On the other hand, β_1^{PA}, may be more appropriate in a public health context where interest is focused on estimating

the likely benefits of increasing or lowering the dose of chemotherapy on the proportion of patients who have an adverse reaction in the population.

Population-averaged and cluster-specific models for continuous y

Marginal models can also be specified for continuous responses. For example, a marginal model for a two-level structure with continuous response y and a single explanatory variable x can be written:

$$\mathrm{E}(\mathrm{y_{ij}}) = \beta_0 + \beta_1 \mathrm{x_{ij}}$$

and, as in the binary response case, we complete the model specification by choosing a correlation structure for the within-group dependencies.

You may therefore wonder why we did not refer to population-averaged models in Module 5! The reason we have not raised this issue before is that population-averaged and cluster-specific coefficients are equal in the continuous response case. It is the nonlinearity of the transformation of π on the left hand side of (7.12) and (7.13) that leads to the difference between generalised linear marginal and random effects models.

C7.3.3 Example: Comparison of marginal and random intercept models fitted to the US election data

The NAES04 data were analysed using single-level, marginal and random intercept logit models. Each model includes the same set of explanatory variables: household annual income (centred about its sample mean), gender, age (centred), and dummy variables for marital status (taking currently married or cohabiting as the reference category). The estimates from the three model specifications are given in Table 7.7. Whatever type of model is fitted, we find statistically significant effects of all variables (based on comparing Z-ratios with a standard normal distribution) and the direction of the effects are the same. As expected from (7.14), however, the cluster-specific coefficients from the random intercept model are all larger in magnitude than the population-averaged coefficients from the marginal model. In this case, the difference between the cluster-specific and population-averaged estimates is small because the between-state variance in the random intercept model is small. When σ_u^2 is larger, the estimated coefficients from a marginal model will tend to be closer to those from a single-level model than to the random intercept estimates.

Table 7.7. Estimates from logit models of voting Bush, US 2004

	Single-level		***Marginal (exchangeable)***		***Random intercept***	
Variable	*Est.*	*SE*	*Est.*	*SE*	*Est.*	*SE*
Constant	0.207	0.029	0.204	0.051	0.204	0.058
Income, mean centred	0.079	0.009	0.094	0.009	0.097	0.009
Female	-0.276	0.035	-0.268	0.034	-0.273	0.035
Age, mean centred	-0.004	0.001	-0.003	0.001	-0.003	0.001

Current marital status (ref.=married or cohab)						
Widowed/divorced)	-0.282	0.045	-0.258	0.045	-0.261	0.046
Never married	-0.641	0.051	-0.570	0.051	-0.579	0.052
Between-state var (σ^2_u)	-	-	-	-	0.107	0.027

C7.4 Predicted Probabilities from a Multilevel Model

In C7.3.2, we discussed the difference in interpretation between population-averaged coefficients from a marginal model and cluster-specific coefficients from a random effects model. In this section, we describe how response probabilities can be calculated from a random effects model (using cluster-specific coefficients) and, in particular, show how to calculate predicted probabilities that have a population-averaged interpretation.

In C7.1.2 we met the following formula for the response probability for individual *i* in group *j*:

$$\pi_{ij} = \frac{\exp(\beta_0+\beta_1 x_{ij}+u_j)}{1+\exp(\beta_0+\beta_1 x_{ij}+u_j)} \qquad (7.15)$$

where, to obtain predicted probabilities, β_0 and β_1 are replaced by their estimates from the fitted model and u_j by the estimated group effect.

Rather than calculate predicted probabilities for each individual in the sample, however, we usually want to make predictions for specific values of *x*, e.g. varying the values of one *x* at a time or for combinations of *x*-values that represent 'typical' individuals (see C6.6.2). For these 'out-of-sample' predictions, we need to decide how to handle the group-level residual u_j. We discuss three approaches here.

Method 1: Substituting u_j = 0, i.e. the mean of the group residuals

We might consider calculating probabilities for specific *x*-values ($x = x^*$ say) holding the group-level residual at its mean of zero. While this is a reasonable strategy, the resultant predictions are not *mean* response probabilities for $x = x^*$. This is because π is a nonlinear function of *u* (see equation (7.15)) and therefore the value of π at the mean of *u* is not equal to the mean of π at $x = x^*$ (averaging across groups).The predicted probability computed at the mean of *u* is in fact the *median* probability at $x = x^*$.[7] However, the mean and median probabilities will be close if the response probability is in the 0.2 to 0.8 range where the logistic (and probit) functions are almost linear. Furthermore, substituting $u = 0$ will give similar predictions to the other approaches described below when the group-level variance σ_u^2 is small.

[7]Consider a set of values on a variable W and suppose we apply a nonlinear transformation to W (e.g. take logarithms or square roots) such that the rank order of the observations is preserved. Then the mean of the transformed values will not in general equal the transformed mean of the original (untransformed) values. For example, the mean of log(W) ≠ log(mean of W). However, the *median* of the transformed values will equal the transformed median of the original values because the rank order of observations is the same after transformation, e.g. median of log(W) = log(median of W). Applying this result to our situation, logit(π) for $x=x^*$ and $u=0$ will not in general equal the mean π for $x=x^*$ across groups. But *u* is normally distributed with mean = median = 0, so that logit(π) for $x=x^*$ and $u=0$ will equal the median π for $x=x^*$ across groups.

Method 2: Integrating out u_j

Method 1 involved substituting a particular value (zero) for the group-level random effect in the formula for the predicted probability. The other two methods we describe lead to predictions that average across different values of *u*, and which therefore have a population-averaged interpretation. The first of these methods involves integrating out the random effects to obtain an expression for the mean response probability for given values of *x* (see Goldstein 2003, p.156). For logit and probit link functions, this integration requires the use of an approximation which can lead to a complex formula, especially when random coefficients are added. We therefore recommend the following simulation approach.

Method 3: Averaging over simulated values of u_j

The aim of Method 2 is to derive a formula for the mean predicted probability for a given set of *x*-values. Mathematically, taking the mean of (7.15) involves performing an integration over the distribution of the random effects *u*. An alternative, more easily implemented approach is to *estimate* the distribution of *u* by generating values for *u* from a normal distribution with mean zero and variance $\hat{\sigma}_u^2$ (the level 2 variance estimate from the fitted model). Based on each generated value, a predicted probability is calculated. The mean probability is then estimated by taking the mean of these predicted probabilities. More formally, the procedure consists of the following steps:

i) Generate M values for random effect *u* from $N(0, \hat{\sigma}_u^2)$ and denote the generated values by $u^{(1)}, u^{(2)} \ldots, u^{(M)}$

ii) For each simulated value ($m = 1,\ldots,M$) compute, for a given value of *x*, $\pi^{(m)} = \frac{\exp(\hat{\beta}_0+\hat{\beta}_1 x+u^{(m)})}{1+\exp(\hat{\beta}_0+\hat{\beta}_1 x+u^{(m)})}$

iii) Calculate the mean of the probabilities computed in (ii):
$$\pi = \frac{1}{M}\sum_{m=1}^{M} \pi^{(m)}$$

iv) Repeat (i)-(iii) for different values of *x*

The computations can be performed quickly in any statistics or spreadsheet package, so M can be large (1000 say). All that is required is a function that generates random numbers from a normal distribution. The procedure has been implemented in MLwiN v2.10, as will be demonstrated in the accompanying practical exercise.

Example: Comparison of predicted probabilities from random intercept and marginal models for US election data

Table 7.8 shows predicted probabilities of voting Bush calculated using the random intercept and marginal model estimates given in Table 7.7. The predictions for the three values of household income are calculated with age, sex and marital status fixed at their sample means (equal to the sample proportions for the sex and marital

status dummies). Similarly, the predictions for sex are calculated with all other variables held constant at their sample means.

Predicted probabilities have been calculated from the random intercept model using two different methods: fixing the state-level random effect at its mean of zero (Method 1) and using a simulation approach with M = 1000 (Method 3). In this case, the probabilities for the two methods are very similar. There are two reasons for this closeness of the two sets of predictions. First, the predicted probabilities are all fairly close to 0.5, the point at which the prediction at *u*=0 (the median) equals the mean prediction (see Figure 7.3). Second, the state-level variance is small at 0.107 (see Table 7.7) so, from equation (7.14), β^{CS} is close to β^{PA}. In longitudinal data, where clusters are individuals, σ_u^2 is typically much larger because measurements on the same individual over time tend to be highly correlated; this leads to large differences between β^{CS} and β^{PA} and, therefore, between predictions computed using Methods 1 and 3.

Probabilities calculated from the random intercept model using Method 3 are almost the same as those from the marginal model. This will generally be the case because both probabilities are population-averaged, i.e. averaged across states.

Table 7.8. Predicted probabilities of voting Bush, US 2004

	Random intercept model		
	Method 1 (u = 0)	*Method 3 (simulated u)*	**Marginal model**
Household income			
Low (band 1)	0.374	0.378	0.377
Medium (band 4)	0.444	0.446	0.445
High (band 9)	0.564	0.564	0.562
Sex			
Male	0.510	0.510	0.510
Female	0.442	0.444	0.444

C7.5 A Two-level Random Slope Model

So far, the multilevel models we have considered have allowed the response probability to vary from group to group by including a group-level random term u_j in the linear predictor of the model. However, this random term affects only the intercept of the model so that the intercept for group *j* is $\beta_0 + u_j$; the effect of each explanatory variable *x* is assumed to be the same in each group. We will now consider random slope models that allow the effect of one or more *x* to vary across groups.

C7.5.1 A random slope logit model

In C5.3 we considered random slope models for a continuous response which involved attaching a random term to one or more explanatory variables. We can do exactly the same in a generalised linear model for a binary response. For example the random intercept logit model given by equation (7.4) can be extended to a random slope model:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \beta_1 x_{ij} + u_{0j} + u_{1j}x_{ij}. \qquad (7.16)$$

As in the continuous response case, we have added a new term $u_{1j}x_{ij}$ to the model and a '0' subscript to the intercept residual. Also as before, the random effects u_{0j} and u_{1j} are assumed to follow normal distributions with zero means, variances σ^2_{u0} and σ^2_{u1} respectively, and covariance σ_{u01}. Because u_{0j} and u_{1j} are allowed to be correlated (i.e. σ_{u01} is not assumed to equal zero), they are said to follow a *bivariate normal* distribution which can be succinctly expressed as

$$u = \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim \mathrm{MVN}(0, \Omega_u)$$

where *MVN* stands for 'multivariate normal' and $\mathbf{0}$ denotes a vector of two zeros, i.e. $0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, and $\Omega_u = \begin{pmatrix} \sigma^2_{u0} & \sigma_{u01} \\ \sigma_{u01} & \sigma^2_{u1} \end{pmatrix}$ is the covariance matrix of the random effects. Note that Ω_u is symmetric, so it is common to give only the diagonal elements (the variances) and the below-diagonal elements (one instance of each of the covariances).

The slope of the linear relationship between *x* and the log-odds that *y* = 1 is $\beta_1 + u_{1j}$ for group *j*. The covariance between the random effects, σ_{u01}, is the covariance between the group intercepts and slopes; we will shortly see how this is interpreted in an analysis of the US election data.

Random intercept probit and clog-log models can be extended to include random slopes in the same way.

C7.5.2 Example: Allowing the relationship between income and voting intentions in the US to vary across states

Table 7.9 shows the results from fitting a random slope logit model for the relationship between the probability of voting Bush and household income, alongside the corresponding random intercept model. For simplicity, we have removed age, sex and marital status from the model, although all were found to be significantly associated with voting intentions.

Table 7.9. Random intercept and slope logit models of voting Bush, US 2004

	Random intercept		*Random slope*	
Parameter	*Est.*	*SE*	*Est.*	*SE*
β_0 (Constant)	-0.099	0.056	-0.087	0.057
β_1 (Income, centred)	0.140	0.008	0.145	0.013
State-level random part				
σ^2_{u0} (intercept variance)	0.125	0.031	0.132	0.032
σ^2_{u1} (slope variance)	-	-	0.003	0.001
σ_{u01} (intercept-slope covariance)	-	-	0.018	0.006

The extension from random intercepts to random slopes has introduced two new parameters to the model - σ^2_{u1} and σ_{u01} - so we can compare the two models by carrying out a test of the null hypothesis that both σ^2_{u1} and σ_{u01} are equal to zero. The test statistic from an approximate Wald test is 9.717 which, when compared to a chi-squared distribution on two degrees of freedom, gives a two-sided p-value of 0.008. Dividing the p-value by 2 to adjust for the fact that σ^2_{u1} cannot be negative (see Section C7.1.3) we obtain 0.004. We therefore conclude that the effect of income does indeed differ across states.

For state *j*, the effect of a one unit increase in income on the log-odds of voting Bush is estimated as 0.145 + $\hat{u}_{1j}$. States showing an above-average positive relationship between income and voting Bush will have $\hat{u}_{1j} > 0$, while states with a below-average positive (or possibly a negative) relationship between income and voting Bush will have $\hat{u}_{1j} < 0$. Figure 7.4 shows the state prediction lines for the relationship between the log-odds of voting Bush and household income. In fact all states have a positive slope so, across all states, it is the richer respondents who are most likely to vote Bush. The intercept variance of 0.132 (see Table 7.9) is interpreted as the between-state variance in the log-odds of voting Bush at the mean household income, while the slope variance of 0.003 is the between-state variance in the effect of income.

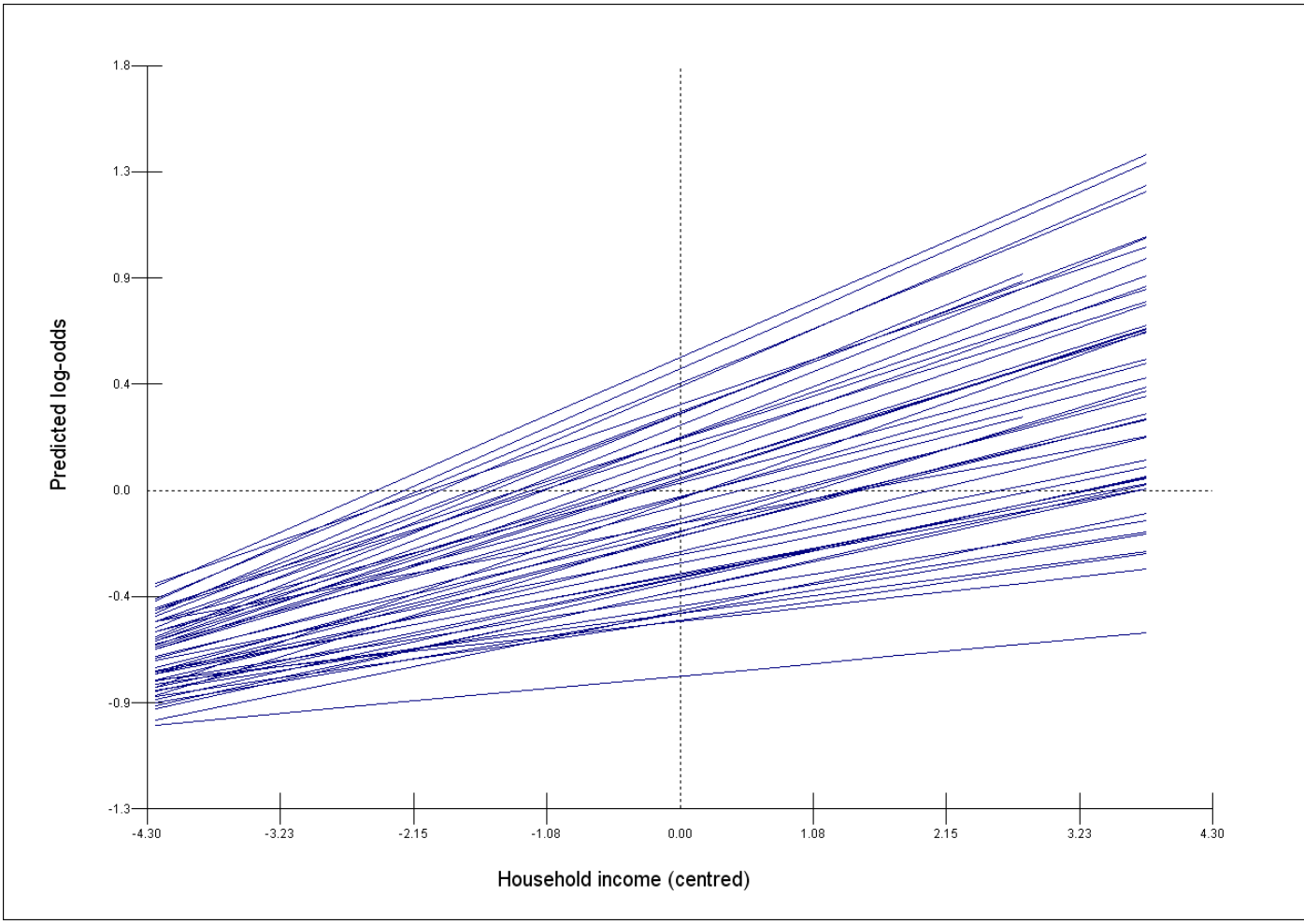

Figure 7.4. Predicted state lines from a random slope logit model of voting Bush, US 2004

The positive covariance between the intercepts and slopes implies that states with a high intercept (higher than average probability of voting Bush, i.e. $u_{0j} > 0$) tend also to have a steep slope (strong positive relationship between voting Bush and income, i.e. $u_{1j} > 0$). This positive covariance, together with the positive estimate for β_1, leads to a 'fanning out' pattern in the prediction lines. The intercept-slope covariance can also be seen in a plot of the intercept residuals $\hat{u}_{0j}$ versus the slope residuals $\hat{u}_{1j}$ (Figure 7.5). The state at the bottom left of the plot with the largest negative intercept and slope residuals is Washington DC; this state has the lowest proportion of Bush voters after controlling for income, and the weakest relationship between voting Bush and income. The prediction line for Washington DC can be seen at the bottom of Figure 7.4, set apart from the other lines by its low intercept and slope. In contrast Montana and Utah, with high proportions of Bush voters and stronger relationships between voting and income, appear at the top right corner of Figure 7.5 and their prediction lines are at the top of Figure 7.4.

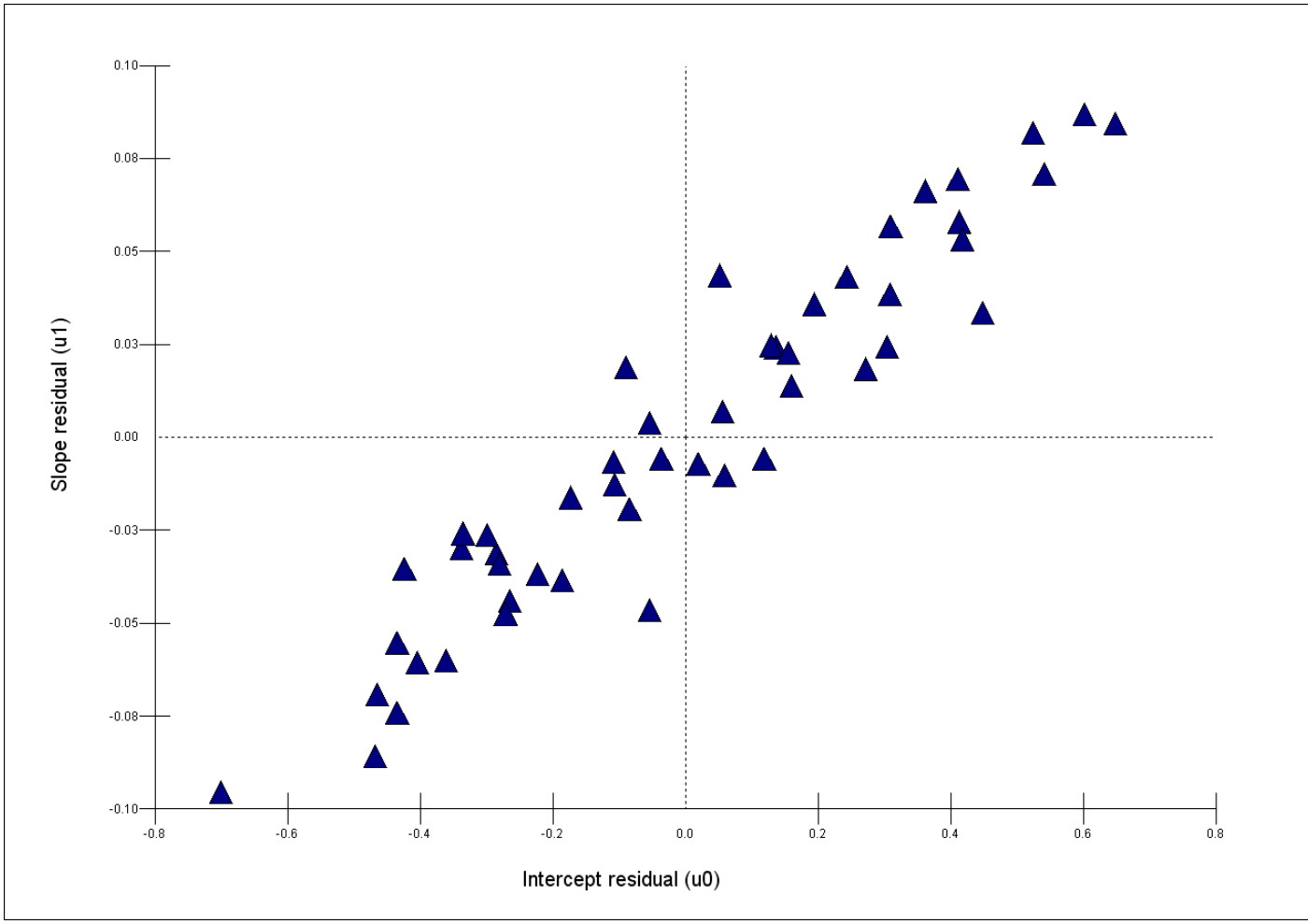

Figure 7.5. Estimated intercept and slope residuals for the relationship between the log-odds of voting Bush and income

The model with a random slope for income implies that the between-state variance depends on income, and the 'fanning-out' pattern in Figure 7.4 suggests that the variance increases with income. In C5.3.5 we gave the formula for the between-group variance in a model with a random slope for an explanatory variable *x* as:

$$\begin{aligned}&\mathrm{var}(u_{0j} + u_{1j}x_{ij}) = \mathrm{var}(u_{0j}) + 2x_{ij}\mathrm{cov}(u_{0j}, u_{1j}) + x_{ij}^2\mathrm{var}(u_{1j}) \\ &= \sigma_{u0}^2 + 2\sigma_{u01}x_{ij} + \sigma_{u1}^2x_{ij}^2\end{aligned} \qquad (7.17)$$

The same formula applies to a random slope model for a binary response. The only difference is that, for a logit model, we are now looking at the variance in the *log-odds that y = 1* rather than the variance in y.

Substituting the estimates from the random slope model in Table 7.9, the between-state variance in the log-odds of voting Bush is estimated as 0.132 + 0.036 Income + 0.003 Income^2. The between-state variance is plotted in Figure 7.6. As anticipated from Figure 7.4, the between-state variance increases with income: between-state differences in the probability of voting Bush are more pronounced among wealthier households.

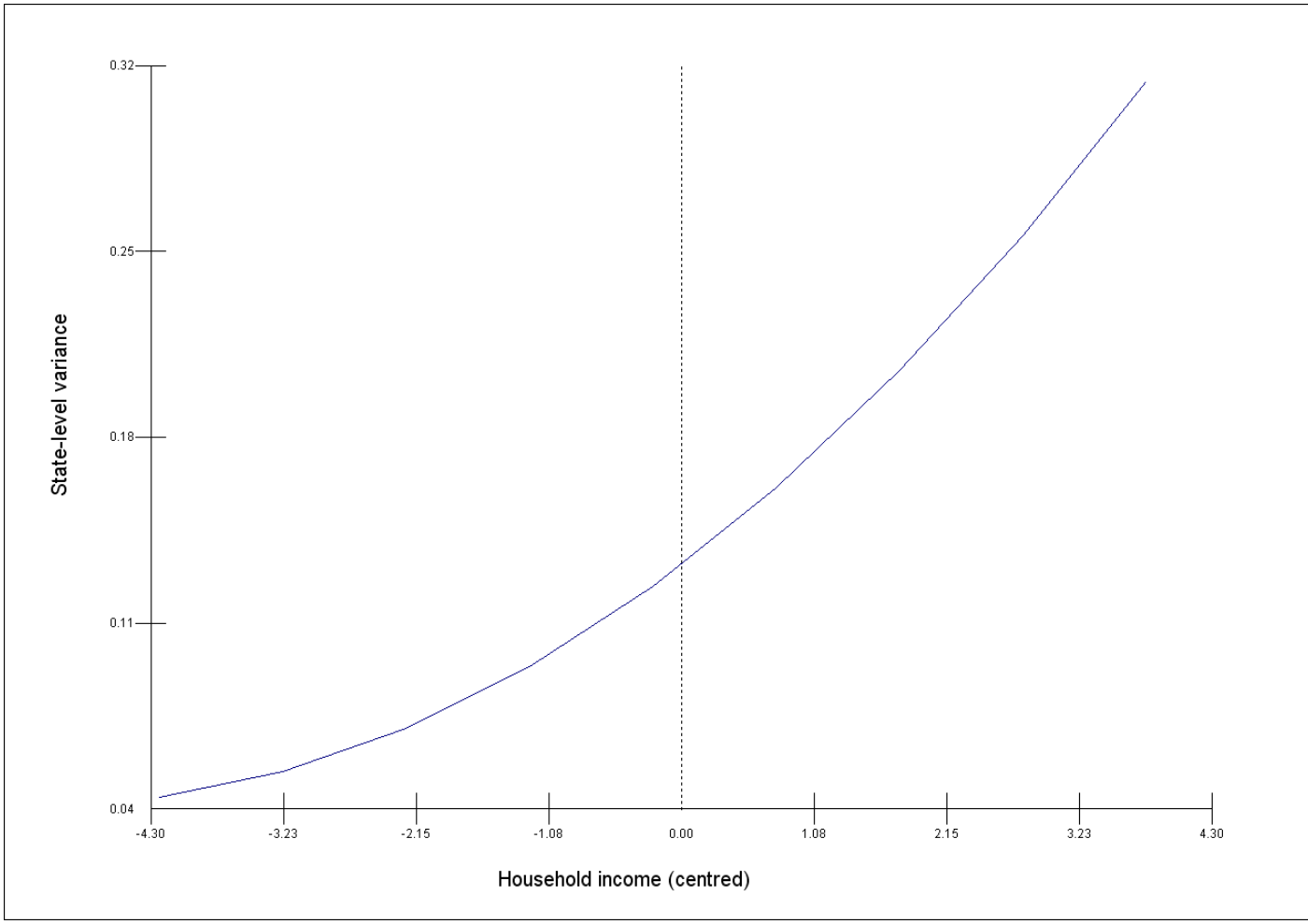

Figure 7.6. Between-state variance in the log-odds of voting Bush as a function of income

C7.5.3 Two random coefficients: Allowing income and urban-rural differentials in voting intentions to vary across states

We can extend the random slope model given by (7.16) to include further explanatory variables and random coefficients for more than one variable. For example, a logit model with two level 1 predictors x_{1ij} and x_{2ij} with all coefficients varying randomly across groups is:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_{0j} + u_{1j} x_{1ij} + u_{2j} x_{2ij} \quad (7.18)$$

where the three group-level random effects follow a trivariate normal distribution:

$$u = \begin{pmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \end{pmatrix} \sim \mathrm{MVN}(0, \Omega_u) \text{ where } \Omega_u = \begin{pmatrix} \sigma^2_{u0} & & \\ \sigma_{u01} & \sigma^2_{u1} & \\ \sigma_{u02} & \sigma_{u12} & \sigma^2_{u2} \end{pmatrix}$$ is the covariance matrix of the random effects.

The model of voting intentions with a random slope for household income (C7.5.2) was extended to include type of region (urban vs. rural) as a predictor, and to allow urban-rural differences to vary across states. The results from fitting a model of the

form (7.18) are presented in Table 7.10. Substituting the estimates of the fixed part parameters β_0, β_1 and β_2 gives the following prediction equation:

$$\log\left(\frac{\hat{\pi}_{ij}}{1-\hat{\pi}_{ij}}\right) = 0.028 + 0.145\ \text{Income}_{ij} - 0.391\ \text{Urban}_{ij} + \hat{u}_{0j} + \hat{u}_{1j}\text{Income}_{ij} + \hat{u}_{2j}\text{Urban}_{ij}$$

where $\hat{\pi}_{ij}$ is the predicted probability of voting Bush for individual *i* in state *j*. The expected effect on the log-odds of voting Bush of a 1-unit increase in income is 0.145 + $\hat{u}_{1j}$ in state *j*, and the expected urban-rural difference is -0.391 + $\hat{u}_{2j}$. The (shrunken) residual estimates $\hat{u}_{1j}$ and $\hat{u}_{2j}$ (together with the intercept residual $\hat{u}_{0j}$) can be computed from the parameter estimates shown in Table 7.10, together with the data.

Table 7.10. Random slope logit model of voting Bush with between-state variance in income and type of region effects, US 2004

Parameter	*Est.*	*SE*
β_0 (Constant)	0.028	0.053
β_1 (Income, centred)	0.145	0.013
β_2 (Urban)	-0.391	0.055
State-level random part		
σ^2_{u0} (intercept variance)	0.105	0.027
σ^2_{u1} (Income slope variance)	0.004	0.002
σ_{u01} (Intercept- income slope covariance)	0.014	0.005
σ^2_{u2}(Urban slope variance)	0.052	0.026
σ_{u02} (Intercept- urban slope covariance)	0.044	0.020
σ_{u12} (Income- urban slope covariance)	0.011	0.005

The random coefficients for Income and Urban imply that the between-state variance depends jointly on Income and type of region. Allowing the urban-rural difference to vary across states introduces three new parameters to the model, given in the last row of the covariance matrix Ω_u: σ_{u02}, σ_{u12} and σ^2_{u2}. The test statistic for the Wald test that all three parameters are equal to zero is 7.15 which, when compared to a chi-squared distribution on 3 degrees of freedom, gives a two-sided p-value of 0.067 and a one-sided p-value of 0.033. We therefore retain the random coefficient for Urban, although we reiterate our earlier point that the Wald test should be interpreted with caution when used to test hypotheses about variances and covariances (see C7.1.3).

We can extend (7.17) to obtain an expression for the between-state variance as a function of Income (x_{1ij}) and Urban (x_{2ij}):

$$\begin{aligned}&\text{var}(u_{0j} + u_{1j}x_{1ij} + u_{2j}x_{2ij}) = \text{var}(u_{0j}) + 2x_{1ij}\text{cov}(u_{0j}, u_{1j}) + x^2_{1ij}\text{var}(u_{1j})\\ &+2x_{2ij}\text{cov}(u_{0j}, u_{2j}) + x^2_{2ij}\text{var}(u_{2j}) + 2x_{1ij}x_{2ij}\text{cov}(u_{1j}, u_{2j})\\ &= \sigma^2_{u0} + 2\sigma_{u01}x_{1ij} + \sigma^2_{u1}x^2_{1ij} + 2\sigma_{u02}x_{2ij} + \sigma^2_{u2}x^2_{2ij} + 2\sigma_{u12}x_{1ij}x_{2ij}\end{aligned}$$

In our application x_{2ij} is binary so the above equation breaks down into two equations: one for urban and another for rural.

For rural areas (x_{2ij} = 0), the between-state variance is

$$\sigma^2_{u0} + 2\sigma_{u01}\,\text{Income}_{ij} + \sigma^2_{u1}\,\text{Income}^2_{ij}$$

which from Table 7.10 is estimated as 0.105 + 0.028 Income + 0.004 Income^2.

For urban areas (x_{2ij} = 1) we have

$$\begin{aligned}&\sigma^2_{u0} + 2\sigma_{u01}\,\text{Income}_{ij} + \sigma^2_{u1}\,\text{Income}^2_{ij} + 2\sigma_{u02} + \sigma^2_{u2} + 2\sigma_{u12}\,\text{Income}_{ij}\\ &= (\sigma^2_{u0} + 2\sigma_{u02} + \sigma^2_{u2}) + (2\sigma_{u01} + 2\sigma_{u12})\,\text{Income}_{ij} + \sigma^2_{u1}\,\text{Income}^2_{ij}\end{aligned}$$

which is estimated as 0.245 + 0.050 Income + 0.004 Income^2.

The variance functions for rural and urban areas are plotted in Figure.7. We can see that, at all levels of income, between-state differences in voting intentions are greater in urban areas. For example, at Income = 0 (the mean) the between-state variance is 0.245 in urban areas and 0.105 in rural areas. We also find that in both urban and rural areas the between-state variance increases with household income, but the increase is sharper in urban areas.

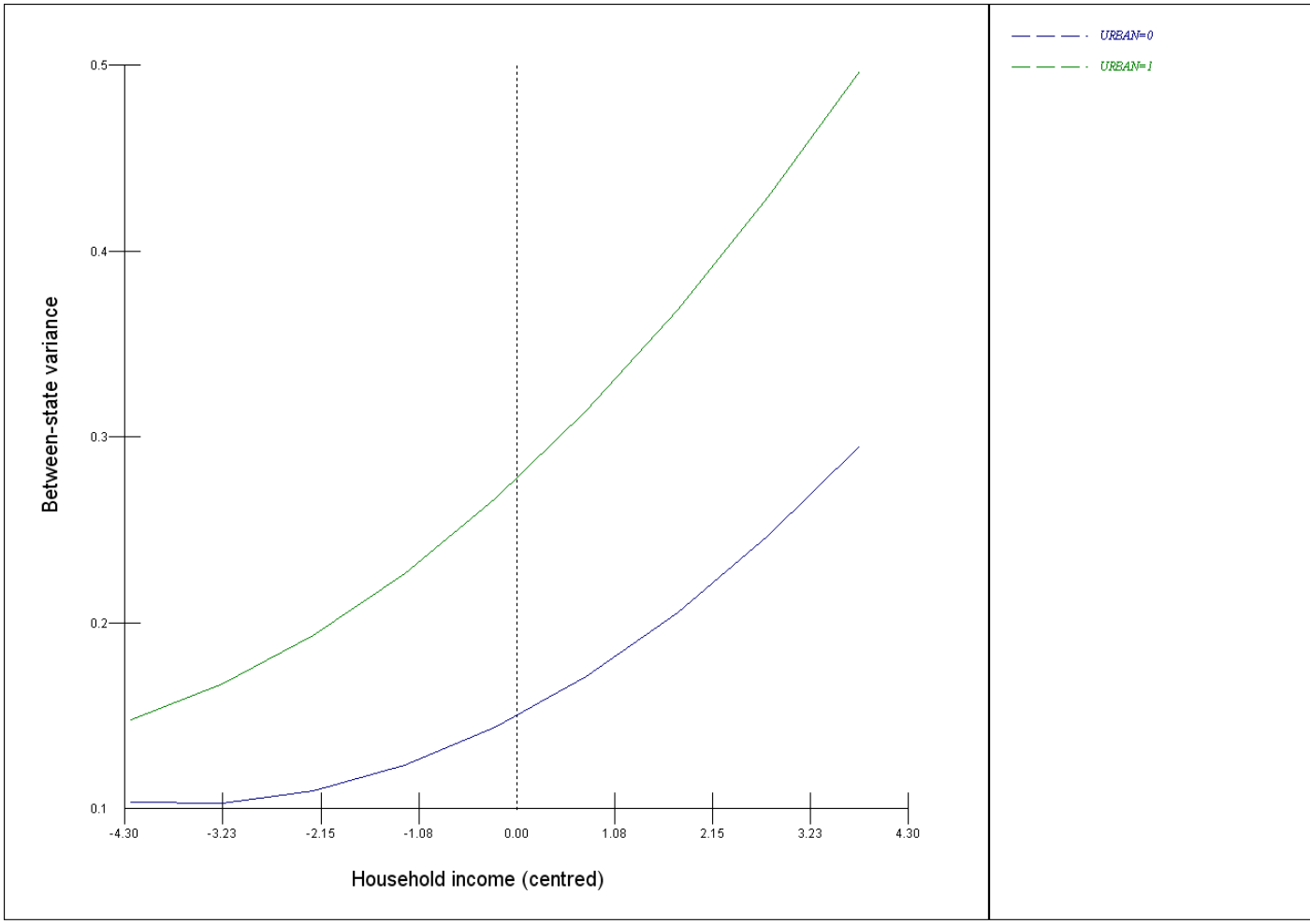

Figure.7. Between-state variance in log-odds of voting Bush as function of income and type of region, US 2004

C7.6 Adding Level 2 Explanatory Variables: Contextual Effects

So far we have considered random intercept and slope models with level 1 explanatory variables. As discussed in Module 5, a particular advantage of multilevel modelling is the ability to explore the effects of group-level (level 2) predictors while simultaneously including random effects to allow for the effects of *unobserved* group-level variables. We referred to the coefficients of level 2 variables as *contextual effects*, and their interactions with level 1 variables as *cross-level interactions* (see C5.4). In this section, we see how contextual effects and cross-level interactions can be included in models for binary responses.

As in the continuous response case, it is particularly important to use a multilevel modelling approach when contextual effects are of interest. The standard errors of coefficients of level 2 variables may be severely underestimated when a single-level model is used.

C7.6.1 A random intercept model with a level 2 explanatory variable

Suppose that we have one explanatory variable defined at level 1, x_{1ij}, and another at level 2, x_{2j}. The random intercept logit model of (7.4) can be extended to include both predictors:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2j} + u_j. \qquad (7.19)$$

In practice, x_{2j} will often be the level 2 mean of a level 1 predictor, e.g. $x_{2j} = \bar{x}_{1j}$. In that case β_2 is the contextual effect of x_1, which is the effect of the group mean of x_1 on an individual's log-odds that is over and above the effect of an individual's own value ofx_1.

We now illustrate the interpretation of contextual effects in our analysis of US voting intentions. For simplicity, we return to the random intercept model fitted at the end of C7.3(with four level 1 predictors - household income, gender, age, and marital status).To this model we add one further level 1 predictor, a dummy variable for frequency of attendance at religious services (coded 1 for weekly or more, and 0 otherwise). We will then add the state-level mean of this variable, i.e. the proportion of respondents who are regular attendees of religious services, to explore whether there is a contextual effect of religiosity that operates over and above any effect of an individual's own religious practice.

Table 7.11 shows the estimates before and after allowing for the state-level proportion of regular attendees at religious services. We find strong and statistically significant individual and contextual effects of religiosity. Respondents who attend services at least weekly are more likely to vote Bush than less regular attendees but, regardless of an individual's own practice, there is a positive effect on an individual's probability of voting Bush of living in a state with a high level of religious practice. We can also see that including the state-level variable leads to a

substantial reduction in the state-level variance: state variation in the proportion of Bush voters is strongly associated with state-level religiosity. While the effect of state-level religiosity is strong, however, the magnitude of its coefficient should be interpreted with caution because it represents a comparison of a state with 100% of respondents attending services at least weekly and a state with 0% attending services as regularly. (We interpret the effect of state-level religiosity for values within the range observed in the sample in a moment.)

Table 7.11. Adding a contextual effect to a random intercept logit model of voting Bush, US 2004

	Model with no contextual effects		***Model with contextual effect of religiosity***	
Parameter	*Est.*	*SE*	*Est.*	*SE*
Constant	-0.010	0.055	-0.879	0.148
Individual-level variables				
Income, mean centred	0.102	0.009	0.102	0.009
Female	-0.331	0.036	-0.332	0.036
Age, mean centred	-0.006	0.001	-0.006	0.001
Marital status (ref.=married or cohabiting)				
Widowed/divorced	-0.207	0.046	-0.207	0.046
Never married	-0.528	0.053	-0.528	0.053
Regular attendance at religious services	0.556	0.037	0.543	0.037
State-level variables				
Proportion attend religious services regularly	-	-	2.151	0.350
Between-state variance (σ_u^2)	0.083	0.022	0.030	0.010

C7.6.2 Cross-level interactions

As in any regression model, we can include interaction effects which allow for the possibility that the effect of one explanatory variable on the outcome depends on the value of another explanatory variable. An interaction between a level 1 variable and a level 2 variable is called a 'cross-level interaction'. We can extend model (7.19) to include an interaction between $x_1$and x_2 as follows:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2j} + \beta_3 x_{1ij} x_{2j} + u_j. \quad (7.20)$$

If we believed, for example, that the effect of an individual's level of religious practice depends on the practice of others living in the same state, we would test for an interaction between an individual's attendance at religious services and the state-level proportion of regular attendees. The null hypothesis in a test for a cross-level interaction is that the coefficient of the interaction variable $x_{1ij}x_{2j}$ equals zero, and a Z-test or Wald test can be used as for any regression coefficient.

In our earlier analyses, we found an individual effect of age and a contextual effect of religiosity on an individual's probability of voting Bush (see Table 7.11). To the extent that regular attendance at religious services is an indicator of closer adherence to traditional values, we might expect a stronger negative effect of age in more conservative states with a high level of religiosity. To explore whether the effect of individual age depends on state-level religiosity, we test for a cross-level interaction between individual age and the state proportion of regular attendees at religious services. Table 7.12 shows the estimates from this cross-level interaction model. The Z-ratio for the interaction coefficient is -0.043/0.013 = -3.31 which is highly significant. Note that the main effect of age is now positive (0.012), but this is the effect of age in a state with no respondents who attend religious services once a week or more (x_2 =0). In fact, the state-level proportion of regular attendees ranges from 0.16 to 0.64 so the effect of age at x_2 = 0 is not very meaningful. We can calculate the effect of age for different values of x_2:

For x_2 = 0.16, the effect of age is 0.012 – (0.043 × 0.16) = 0.005
For x_2 = 0.30, the effect of age is 0.012 – (0.043 × 0.30) = -0.0009
For x_2 = 0.64, the effect of age is 0.012 – (0.043 × 0.64) = -0.016

Thus the age effect is weakly positive for the least religious states, and becomes less strongly positive and then more strongly negative as state-level religiosity increases. The difference between young and old respondents in their probability of voting Bush is greatest in the most religious states. The cross-level interaction effect is illustrated graphically in Figure 8 which shows the relationship between the (population-averaged) predicted probability of voting Bush and individual age for state proportions of regular attendees at religious services of 0.15, 0.3, 0.45 and 0.6.

Table 7.12. Estimates from a random intercept logit model of voting Bush with a cross-level interaction, US 2004

Variable	*Est.*	*SE*
Constant	-0.886	0.147
Individual-level variables		
Income, mean centred	0.101	0.009
Female	-0.331	0.036
Age, mean centred	0.012	0.005
Current marital status (ref=married or cohabiting)		
Widowed/divorced	-0.207	0.046
Never married	-0.524	0.053
Regular attendance at religious services	0.543	0.037
State-level variables		
Prop. attend religious services regularly	4.206	0.716
Cross-level interaction		
Age × state-level religious attendance	-0.043	0.013
Between-state variance (σ_u^2)	0.029	0.010

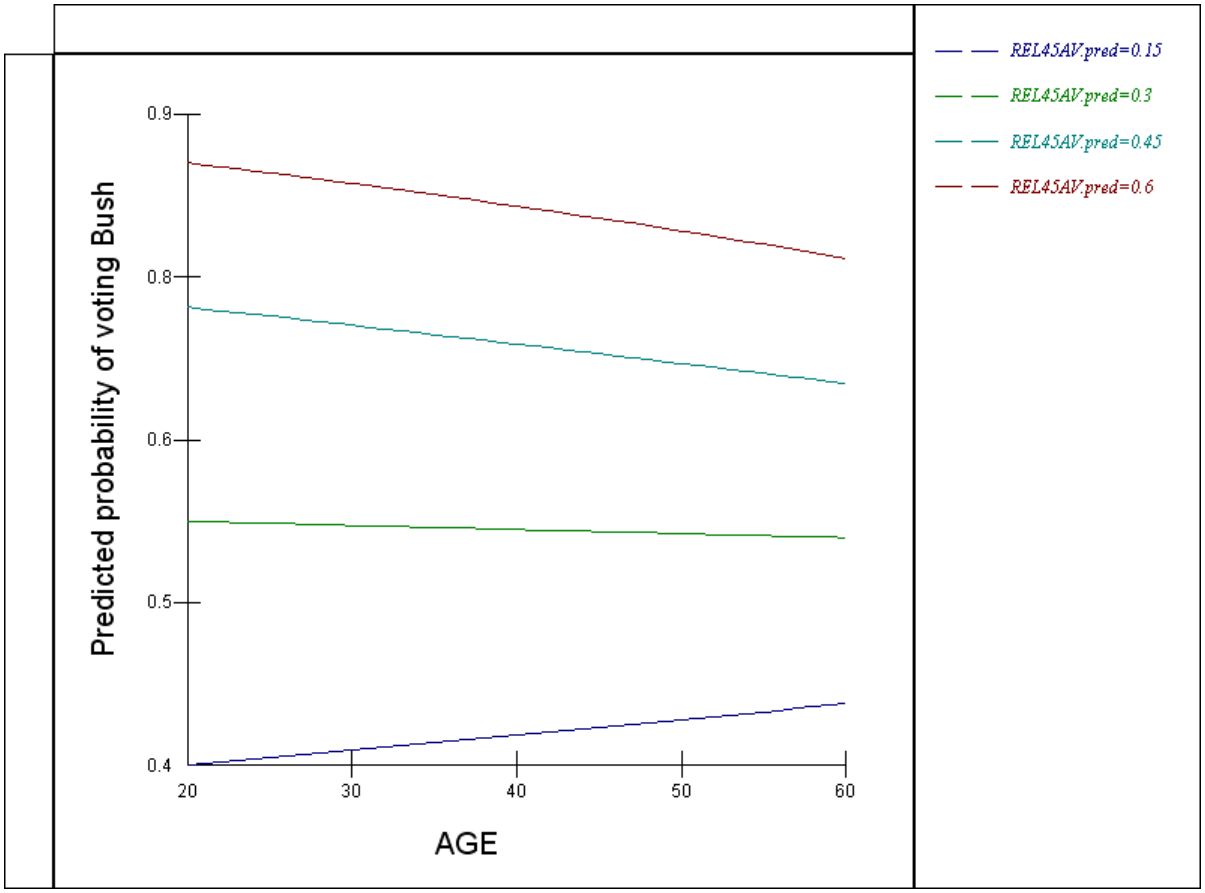

Figure 8. Predicted probabilities of voting Bush by individual age and state-level religiosity

C7.7 Estimation of Binary Response Models

Multilevel models for continuous responses are usually fitted using maximum likelihood procedures. For example, maximum likelihood estimates are obtained in MLwiN using the iterative generalised least squares (IGLS) algorithm, a brief description of which is given in the Technical Appendix. In the case of binary responses, however, maximum likelihood procedures are highly computer-intensive for all but the relatively simple variance components model. As such, a number of alternative estimation procedures for binary responses have been proposed in the statistics literature; but a problem for the applied researcher is how to choose between them, especially when in some situations the different procedures can lead to quite different results. The Technical Appendix provides a very brief overview of some of the estimation procedures implemented in mainstream statistical software. In this section we show simulation results on comparisons of different methods, and offer some practical guidance on choosing between them.

We consider the following widely used estimation procedures, all implemented in mainstream as well as specialist statistical software packages:

- direct *maximum likelihood* via numerical integration (implemented in, for example, SAS, Stata, MIXOR and aML)
- *quasi-likelihood methods* (MLwiN and HLM), including marginal and penalised quasi-likelihood (MQL and PQL)
- *Markov chain Monte Carlo (MCMC) methods* (WinBUGS and MLwiN) where Gibbs sampling and Metropolis Hastings are the most widely used methods

C7.7.1 Comparison of estimation procedures

Rodríguez and Goldman (2001) conducted a simulation study to compare results from fitting multilevel logit models using different estimation procedures. Their simulated data was based on the structure of a Guatemalan dataset with 2449 births (at level 1) from 1558 mothers (level 2) living in 161 communities (level 3). The simulated data included one predictor at each level, and the binary outcome (child immunisation status) was generated under a three-level random intercept model with the coefficient of each predictor equal to 1, and standard deviations of the family and community random effects also set to 1. A total of 100 datasets were generated, and each was analysed with a 3-level logit model using various quasi-likelihood procedures: first and second order marginal quasi-likelihood (MQL1 and MQL2) and second order penalised quasi-likelihood (PQL2) methods. The 100 sets of parameter estimates were then averaged to give the results shown in Table 7.13. In each simulated dataset, the mean number of children per woman is 1.57 and a high proportion of women had only one child, so the family clusters are small. Quasi-likelihood methods are expected to perform poorly in this type of situation, and this is borne out in the simulation results. The estimates from all methods are biased downwards, but PQL2 is a considerable improvement over MQL.

Table 7.13. Results from Rodríguez and Goldman's simulation study†

Parameter	*True value*	*MQL1*	*MQL2*	*PQL2*
Child-level variable	1	0.74	0.85	0.96
Family-level variable	1	0.74	0.86	0.96
Community-level variable	1	0.77	0.91	0.96
Random effect standard deviations				
Family	1	0.10	0.28	0.73
Community	1	0.73	0.76	0.93

†Source: Table 1 from Rodríguez and Goldman (2001)

Rodríguez and Goldman then analysed real Guatemalan data on child immunisation using several methods: MQL and PQL (both 1st and 2nd order), PQL1 with bias correction using an iterated bootstrap procedure (PQL1-B), maximum likelihood (via numerical integration) and MCMC (Gibbs sampling). The standard deviations of the family and community-level random effects from their analysis are given in Table 7.14. (The MQL and PQL1 estimates are not shown but, as in the simulation study, they are biased downwards and much smaller than the estimates obtained from the other methods.) The estimates of the random effect deviations are fairly similar for PQL1-B, maximum likelihood and MCMC.[8]

Table 7.14. Selected results from Rodríguez and Goldman's analysis of the Guatemalan child immunisation dataset†

Random effect standard deviations	*PQL2*	*PQL1-B*	*ML*	*MCMC*
Family	1.75	2.69	2.32	2.60
Community	0.84	1.06	1.02	1.13

†Source: Table 2 from Rodríguez and Goldman (2001)

C7.7.2 Some practical guidelines on the choice between estimation procedures

- *Maximum likelihood via numerical integration* is generally the preferred method for relatively simple random effects models. However, estimation times may be lengthy for models with several random effects fitted to large datasets, especially if the random effects are correlated (e.g. in a random coefficient model).

- *Quasi-likelihood methods* are quick and useful for initial model screening, but a bias correction method should be used when applied to datasets with small cluster sizes (e.g. individuals within households or repeated measures). At a

[8] Rodríguez and Goldman also included a number of explanatory variables (at the child, family and community levels) and compared coefficients for the different methods. The results for PQL1-B, maximum likelihood and MCMC are very similar.

minimum, PQL2 should be used and the final results compared with results from an alternative estimation procedure such as MCMC.

- *MCMC methods* are flexible and can be used for estimating complex multilevel models for binary responses and a range of other response types. However, although they are becoming increasingly computationally feasible, estimation times are still long when applied to large datasets. A further advantage of MCMC methods is that, being Bayesian, we can introduce informative prior distributions where this is felt desirable.

References

Goldstein H. (2003) *Multilevel Statistical Models*. 3rd edn. Arnold, London.

Hedeker D. and Gibbons R.D. (2006) *Longitudinal Data Analysis*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Rodríguez G. and Goldman N. (2001) Improved estimation procedures for multilevel models with binary response: a case-study. *Journal of the Royal Statistical Society, Series A*, 164, 339-355.

Snijders T.A.B. and Bosker R.J. (1999) *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Sage Publications Ltd, London.