Module 7: Multilevel Models for Binary Responses

Concepts

Fiona Steele[1]
Centre for Multilevel Modelling

Pre-requisites box
• 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

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

For a binary response y_{ij}, we have $E(y_{ij} \mid x_{ij}, u_j) = \pi_{ij} = \Pr(y_{ij} = 1)$ 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 (cloglog) 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

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

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 $\text{var}(u_j) = \sigma_u^2$, 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.

[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.

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:

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

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

This document is only the first few pages of the full version.

To see the complete document please go to learning materials and register:

http://www.cmm.bris.ac.uk/lemma

The course is completely free. We ask for a few details about yourself for our research purposes only. We will not give any details to any other organisation unless it is with your express permission.