Module 10: Single-level and Multilevel Models for Nominal Responses Concepts

Fiona Steele
Centre for Multilevel Modelling

Pre-requisites

- Modules 5, 6 and 7

Contents

Introduction

In Module 6 we saw how multiple regression models can be generalised to handle binary responses, and in Module 7 these models were extended for the analysis of binary data with a two-level hierarchical structure. Module 9 considered single-level and multilevel models for categorical responses with more than two categories, where the numeric codes assigned to categories imply an ordering. Examples of ordinal variables include Likert scale items where respondents are asked to indicate their strength of agreement with a statement, and exam grades. In this module we look at models for nominal (or unordered) categorical responses, where the numeric codes assigned to categories are simply labels and serve only to distinguish between categories (see C1.3.8 for a classification scheme for variables).

Examples of nominal responses include political party preferences (e.g. Labour, Conservative, Liberal Democrat, other in the UK), mode of transport and brand preference. Aggregating such variables to a binary response not only wastes potentially important information, but may result in misleading conclusions if predictors have different effects for different categories. For example, the choice between driving to work or using public transport may depend on the availability of free car-parking, while the choice between driving and walking is likely to depend strongly on the distance between home and work. Fortunately, multinomial regression methods have been developed that allow such distinctions between categories of a nominal response, and these have been extended to handle multilevel data structures.

In this module, we begin by describing multinomial logit models for single-level nominal responses. As the coefficients of multinomial models can be difficult to interpret, we pay particular attention to calculating predicted response probabilities to aid interpretation. We then consider multilevel multinomial logit models for two-level structures. We shall see that models for nominal responses are direct extensions of the models for binary responses described in Modules 6 and 7. The same generalisations of the basic multilevel model – for example, random slopes and contextual effects – are possible for nominal responses. We end with a discussion of conditional logit models which are used when the effects of characteristics of the different response alternatives are of interest. For example, the choice between driving and using public transport may depend on their relative costs to an individual, according to where the individual lives and the travel time for each option.

Introduction to the Example Dataset

Our main example dataset for this module comes from the 2008 National Travel Survey (NTS)[1]. The 2008 NTS is one of a series of annual cross-sectional household

[1]Department for Transport, *National Travel Survey, 2002-2008* [computer file]. *5th edition*. Colchester, Essex: UK Data Archive [distributor], June 2010. SN: 5340. The data are free to download after registration from http://www.data-archive.ac.uk/

surveys, designed to provide regular data on personal travel in Great Britain. We will use data from personal face-to-face interviews (the survey also includes travel diaries), and restrict the sample to household members who were aged 16 or older.

The response variable for the analysis is the mode of transport used to travel to work, which has been grouped into three categories:

Code	Label
1	Car /motorcycle
2	Bicycle or walking
3	Public transport

We consider three individual-level characteristics as explanatory variables (all categorical):

- *Gender*
- *Age* (16-19, 20-29, 30-39, 40-49, 50-59 years)
- *Employed part-time* (versus full-time)

The survey is based on a stratified two-stage random probability sample of private households in Great Britain. The primary sampling units (PSUs) at the first stage of sampling are postcode sectors. At the second stage, a sample of households was drawn from the selected PSUs.[2] We will ignore the household level in this module, and treat the data as a two-level structure with individuals at level 1 and PSUs at level 2.

We consider one PSU-level explanatory variable:

- *Type of area* (London boroughs, metropolitan built-up areas, other urban areas over 250,000 population, urban 25,000-250,000 population, urban 10,000-25,000 population, urban 3000-10,000 population, rural)

After excluding a small number of individuals with missing data on at least one of the variables, the analysis file contains 8,512 individuals nested within 683 PSUs.

Note that the same dataset was analysed in Module 9 for an ordinal response (frequency of walking). In this module, the analysis sample has been restricted to employed respondents aged less than 60 because means of travel to work was only asked of this group.

[2] See Anderson, Christophersen, Pickering, Southwood and Tipping (2009) *National Travel Survey 2008 Technical Report*. Prepared for the Department of Transport. This report and other documentation can be downloaded with the dataset from http://www.data-archive.ac.uk/

C10.1 Multinomial Logit Model for Single-Level Data

In this module we focus on multinomial logit models, the most common approach for the analysis of nominal responses. Another model for nominal responses, the conditional logit model, is discussed in the final section C10.5.

C10.1.1 The multinomial logit model

Consider response variable y which takes values 1, 2, . . ., C.

We define *response probabilities* for each category k as

$\Pr(y = k) = \pi_k$

where $\pi_1 + \pi_2 + \ldots \pi_C = 1$.

As for binary and ordered logit models, one of the response categories is chosen as the reference. We then model the log-odds of being in one of the remaining categories rather than the reference category. If we take the first category as the reference, for example, we model the log-odds of being in category k $(k = 2, \ldots, C)$ rather than category 1.

We begin be considering models for a single-level nominal response. Suppose we have one continuous or binary explanatory variable x, then the model for the contrast between response category k and the reference category 1 for individual i $(i = 1, \ldots, n)$ can be written

$$\log\left(\frac{\pi_{ki}}{\pi_{1i}}\right) = \beta_{0k} + \beta_{1k}x_i, \quad k = 2, \ldots, C \qquad (10.1)$$

Equation (10.1) consists of $C - 1$ contrasts or sub-equations, one for each category apart from the reference, where β_{0k} is the intercept and β_{1k} the effect of x for the contrast of category k versus category 1.

Before discussing interpretation of the multinomial logit model, we note that the binary logit described in Module 6 is a special case of (10.1). To see this, suppose that the response y_i is binary but coded 1 and 2 (rather than the usual 0 and 1). Taking the first category as the reference (now coded 1 rather than 0) equation (10.1) reduces to a single contrast:

$$\log\left(\frac{\pi_{2i}}{\pi_{1i}}\right) = \log\left(\frac{\pi_{2i}}{1-\pi_{2i}}\right) = \beta_{02} + \beta_{12}x_i,$$

where π_{2i} is the binary response probability.

Remarks

- The multinomial logit model given by (10.1) has the same predictor x in each equation. This restriction can be relaxed to allow a predictor to affect a subset of contrasts. In some software packages, it is possible to directly specify the contrast(s) for which a particular predictor should be included. In other packages, a predictor is removed from a contrast by constraining its coefficient to equal zero.

- The equations in (10.1) are estimated simultaneously, but an approximation to the multinomial logit model is obtained by estimating a series of binary logit models on subsets of the data. For example, the contrast of category 2 versus 1 may be approximated by selecting respondents with $y = 1$ or $y = 2$ and estimating a simple logit model for a new binary response distinguishing these two categories (coded 1 when $y = 2$ and 0 when $y = 1$).[3] However, this approach does not extend to the multilevel case where we will typically wish to allow for correlations between random effects for the different contrasts.

C10.1.2 Interpretation of coefficients and predicted probabilities

The intercept β_{0k} for contrast k is the log of the probability of being in category k relative to the probability of being in category 1 when $x = 0$, and its exponent $\exp(\beta_{0k})$ is the ratio of the probability of being in category k to the probability of being in category 1. The left-hand side of equation (10.1) is commonly referred to as the log-odds of being in category k rather than category 1, and we will refer to it as such as a shorthand even though we are really modelling the ratio of two probabilities.[4] However, it is incorrect to refer to $\log(\pi_{ki}/\pi_{1i})$ as simply the odds of being in category k (as we would for a binary response); if we do not explicitly refer to the reference category, the odds are $\log(\pi_{ki}/(1-\pi_{ki}))$. This is an important difference between the binary logit model and the multinomial logit model for a multi-category response which has implications for the interpretation of coefficients from a multinomial model (as discussed below).

The coefficient of x for contrast k, β_{1k}, is the effect of a 1-unit increase in x on the log-odds of being in category k rather than category 1. As in the binary response case, we can interpret $\exp(\beta_{1k})$ as an odds ratio, comparing the odds of being in category k rather than category 1 for two randomly selected individuals whose x values differ by 1 unit.

As you can tell from the above, interpretation of the coefficients of a multinomial logit model (and the associated odds ratios) is rather awkward! In a binary logit model, the coefficients are the effects of predictors on being in one of the

[3] This approximation was proposed by Begg, C.B. and Gray, R. (1984) "Calculation of polychotomous logistic regression parameters using individualized regressions". *Biometrika* **71**, 11-18.

[4] Exponentiated coefficients from a multinomial logit model are more accurately described as *relative risk ratios*, but this terminology is less commonly used than *odds ratio*.

response categories rather than the other, but in the multinomial generalisation we could have many pairwise contrasts to consider. It would be much easier to interpret the effects of a predictor on each response category, rather than on a contrast between two categories. Fortunately, we can calculate predicted response probabilities from the estimated coefficients for whatever values of x we choose.

Equation (10.1) can be rearranged to give the following expressions for the response probabilities:

$$\pi_{ki} = \frac{\exp(\beta_{0k} + \beta_{1k}x_i)}{1 + \sum_{l=2}^{C}\exp(\beta_{0l} + \beta_{1l}x_i)}, \quad k = 2, \ldots, C \qquad (10.2)$$

with the probability for the reference category calculated by subtraction:

$$\pi_{1i} = 1 - \sum_{l=2}^{C}\pi_{li} = \frac{1}{1 + \sum_{l=2}^{C}\exp(\beta_{0l} + \beta_{1l}x_i)} \qquad (10.3)$$

Predicted response probabilities are calculated by 'plugging in' the estimates for β_{0k} and β_{1k} from the fitted model and applying (10.2) and (10.3) for selected values of x (some examples will be given in C10.2).

Retherford and Choe (1993, p.153)[5] note that coefficients (or odds ratios) are not only difficult to interpret, but may even be misleading because the sign of β_{1k} may not reflect the direction of the effect of x on either of the response probabilities being compared (π_k and π_1). To illustrate the problem, suppose we fit a multinomial logit model to a three-category response taking category 1 as the reference, and including a single binary predictor x. We consider two scenarios where the coefficient of x for the contrast of response categories 2 and 1, β_{12} in equation (10.2), does not reflect the effect of x on the response probabilities for these categories.

In Table 10.1 the probabilities for categories 1 and 2 (π_1 and π_2) are both lower for $x = 1$ than for $x = 0$, so we would say that there is a negative association between being in categories 1 or 2 of the response and x. However, the ratio of π_2 to π_1 is constant across values of x, so that $\exp(\beta_{12}) = 1$ which implies $\beta_{12} = 0$. Interpreting the coefficients of x, we might be tempted to incorrectly conclude that there is no relationship between x and being in response category 2. The correct interpretation of β_{12} is that the probability of being in category 2 *rather than category 1* does not depend on x.

[5] Retherford, R. D., & Choe, M. K. (1993). *Statistical Models for Causal Analysis*. New York: Wiley.

Module 11: Three-Level Multilevel Models

George Leckie
Centre for Multilevel Modelling

Pre-requisites

- Modules 1-5

Contents

If you find this module helpful and wish to cite it in your research, please use the following citation:

Leckie, G. (2013). Three-Level Multilevel Models - Concepts. LEMMA VLE Module 11, 1-47.
http://www.bristol.ac.uk/cmm/learning/course.html

Address for correspondence:

George Leckie
Centre for Multilevel Modelling
University of Bristol
2 Priory Road
Bristol, BS8 1TX
UK

g.leckie@bristol.ac.uk

What are Three-Level Multilevel Models?

In the previous modules we illustrated two-level multilevel models for analysing two-level data structures where units (level 1) are nested within groups (or clusters) (level 2). When the groups are themselves nested within supergroups (or superclusters) (level 3), the data form a three-level hierarchy and three-level models can be fitted to account for the additional level. Examples of three-level data structures include: students (level 1) nested within classes (level 2) nested within schools (level 3); voters nested within counties nested within states; and patients nested within doctors nested within clinics. In this module, we describe three-level data structures and multilevel models which can be used to analyse them. Of course, there is nothing to stop data structures being even more complex and having four or more levels and we shall also consider examples of such data structures in this module. Many further examples of three- and four-level data structures are described in C4.2 and C4.3 of Module 4, respectively.

It is important to incorporate three-level structures in to our models when they arise in the data and lead the higher level clusters to differ substantially from one another on the response variable under study. Naively fitting two-level models to three-level data will lead us to misattribute response variation to the two included levels (van Landeghem et al., 2005; Moerbeek, 2004; van den Noortgate et al., 2005; Tranmer and Steele, 2001). This in turn may lead us to draw misleading conclusions about the relative importance of different sources of influence on the response. For example, fitting a students-within-classes two-level model of student attainment and ignoring the fact that classrooms are further nested within schools will likely lead us to overstate the importance of classrooms as a source of variation in student attainment. That is, much of the variation that we attribute to classrooms may be driven by school-to-school differences in attainment. Our naïve analysis would therefore overstate the importance of classrooms on student attainment and would ignore the role of schools (i.e. school policies, practices, context and compositional effects). Furthermore, by incorrectly modelling the dependency in the data we will likely obtain biased standard errors for the predictor variables, particularly those measured at higher levels. We therefore run the risk of making incorrect inferences and drawing misleading conclusions about the relationships being studied. For example, including school-level predictor variables in our students-within-classes two-level model, but ignoring school as a level in the model will typically lead us to severely underestimate the standard errors on these school-level variables. When we then go on to test the significance of these variables, we will run the risk of making type 1 errors of inference.

Introduction to the Example Dataset

In educational research, there is considerable interest in measuring the effects that schools have on students' educational achievements. Measuring the effects that schools have on their students is after all a necessary first step to learning how schools' policies and practices combine to generate differences between schools. Governments are also often interested in measuring school effects, typically for school accountability purposes, but often to also provide parents with information to help guide school choice. However, in nearly all education systems, there are substantial differences between schools in their students' attainments at intake (i.e. when students first arrive at their schools). For the purposes of researching the effects of schools' policies and practices, holding schools accountable, or informing school choice, schools should not be compared simply in terms of their average exam results as these differences will, at least in part, be driven by these initial differences.

Traditional studies of school effects attempt to measure the 'true' effects that schools have on their students by fitting two-level students-within-schools multilevel models to students' exam scores where covariate adjustments are made for students' initial scores, and typically for a range of other student background characteristics. The school-level residuals from these models are then argued to measure the effects that schools have on their students having adjusted for the non random selection of students into schools. These effects are interpreted as measuring the influences schools have on their students' academic progress (improvement or change in attainment) while they attend their schools. In school effectiveness research these influences are referred to as 'value-added' effects.

In terms of studying students' academic progress, there are many other potential sources of clustering or influence which may also be important determinants of student progress. For example, where data contain multiple academic cohorts of students, we can think of schools as potentially having different effects in different academic cohorts. This leads students from the same school-cohort to appear more alike than students from different school-cohorts. The data are then three-level with students (level 1) nested within school-cohorts (level 2) nested within schools (level 3). In this module, we shall introduce three-level multilevel models to explore such data. In particular, we shall focus on the stability of school effects over time by examining the extent to which school effects change from cohort to cohort.

We shall then go on to consider the further nesting of schools within administrative educational regions referred to as local authorities (LAs) (level 4).[1] In England, secondary schools are organised into 150 LAs. Traditionally, LAs controlled the distribution of government funds across schools, co-ordinated school admissions, and were the direct employers of all teachers and staff in many schools. While over the last few years there has been a reduction of LAs' powers, one might still expect to identify LA effects in the data. If nothing else, we would expect LA

[1] LAs correspond to school districts in the U.S.

effects to pick up geographic variation in student attainment that exists across England.

We shall use data from England's National Pupil Database (NPD), a census of all students in state (i.e. government funded) schools in England. The data are provided by the Department for Education (http://www.education.gov.uk). The NPD records students' academic attainments and a limited number of background characteristics. We focus on three consecutive academic cohorts of students who sat their General Certificate of Secondary Education (GCSE) examinations (age 16 years) in London schools in 2008, 2009 and 2010, respectively. These students sat their Key Stage 2 (KS2) examinations (age 11 years) five years earlier in 2003, 2004 and 2005, respectively.[2] [3]

Table 11.1 presents the number of units at each level of this data hierarchy.

Table 11.1 Number of units at each level of the data hierarchy

Level number	Level	Number of units
4	LAs	32
3	Schools	427
2	School-cohorts	1,232
1	Students	189,940

Thus, there are 32 LAs at level 4, 427 schools at level 3, 1,232 school-cohorts at level 2 and 189,940 students at level 1 of the data hierarchy. At this point it is helpful to explicitly define 'cohort' and 'school-cohort'. When we say 'cohort' we are referring to the three academic cohorts in the data: 2008, 2009 and 2010. When we refer to 'school-cohorts' we are referring to the 1,232 groups, or school-by-cohort combinations of students, in the data which are formed by crossing the 427 schools by the three cohorts. The number of schools and students present in the data for each cohort are as follows. In 2008 there were 412 schools and 63,208 students. In 2009 there were 410 schools and 63,072 students. In 2010 there were 410 schools and 63,660 students.[4] Three hundred and ninety five schools had all three cohorts represented in the data, 15 schools had only two of the three cohorts, while a further 17 schools had only one of the cohorts present. The 32 schools which were not present for one or more cohorts reflect the opening of new schools and the closing of old schools.

The response variable for all our analyses is a continuous point score summarising students' overall attainment in their GCSE examinations.[5] To ease the

[2] GCSE examinations are taken in the last year of secondary schooling. Successful GCSE results are often a requirement for taking A-level examinations (age 18 years) which in turn are a common type of university entrance determinant. For those who leave school at 16 years of age, GCSE results are their main job market qualification.
[3] KS2 examinations are taken in the last year of primary schooling.
[4] The 2010 cohort of 410 schools and 63,660 students will provide the example dataset in Modules 11 and 12.
[5] Specifically, the response variable is the student's capped 'best 8' total point score at GCSE with an additional bonus for attainment in each of English and Mathematics, and is the same measure as

interpretation of this variable, and so that the residuals at each level better approximate the normality assumptions of the models, we transform it to a standard normal score which has the property of being more normally distributed with mean zero and variance one.[6] This transformation allows the effects of the covariates in our multilevel models to be interpreted in terms of standard deviation units of the response. As our focus is on the stability of school effects across cohorts and not on any overall, London-wide, trend in student attainment over time, we carry out this transformation separately for each cohort. Put differently, in this analysis we are interested in the relative performance of schools to one another; we are not interested in the average absolute performance of schools.

We consider eight student-level predictor variables

- Attainment at age 11 (average point score across English, maths and science) (transformed to a standard normal score)
- Female (0 = male; 1 = female)
- Age (ranges from 0 to 1 where higher values correspond to older children; specifically, 0 corresponds to the youngest child in the data, born on the last day of the academic year, while 1 corresponds to the oldest child in the data, born on the first day of the academic year)
- Eligible for free school meals (FSM) (0 = no FSM; 1 = FSM)
- Special education needs (SEN) (0 = no SEN; 1 = SEN)
- English as an additional language (EAL) (0 = no EAL; 1 = EAL)
- Ethnicity (1 = White; 2 = Mixed; 3 = Asian; 4 = Black; 5 = Chinese; 6 = Other)
- Index of deprivation affecting children index (IDACI) a measure of residential neighbourhood social deprivation (transformed to a standard normal score)

and one school-cohort-level variable

- Cohort (1 = 2008; 2 = 2009; 3 = 2010)

---

that published in Government school performance tables (see http://www.education.gov.uk/performancetables).

[6] The transformation is carried out by first ranking the N students by their original scores. The standard normal score for the ith ranked student in the data is then $\Phi^{-1}\{(i-0.5)/\mathrm{N}\}$, where Φ^{-1} denotes the inverse of the standard normal cumulative distribution function. This transformation is order preserving and students with the same original scores will also be tied in terms of their standard normal scores.

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.