Module 6: Regression Models for Binary Responses Concepts

Fiona Steele[1]
Centre for Multilevel Modelling

Pre-requisites

- Modules 1-3

Contents

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

All of the sections within this module have online quizzes for you to test your understanding. To find the quizzes:

From within the LEMMA learning environment

- Go down to the section for **Module 6: Regression Models for Binary Responses Concepts**
- **Click "6.1 Preliminaries: Mean and Variance of Binary Data " to open Lesson 6.1**
- Click Q 1 to open the first question

Most of the sections within this module have practicals so you can learn how to perform this kind of analysis in MLwiN or other software packages. To find the practicals:

From within the LEMMA learning environment

- Go down to the section for **Module 6: Regression Models for Binary Responses Concepts**

Then either

- Click "6.1 Preliminaries: Mean and Variance of Binary Data " to open Lesson 6.1
- Click MLwiN practical

Or

- Click Print all Module 6 MLwiN Practicals

Introduction

In Module 3 we considered multiple linear regression models for the relationship between a continuous response variable (*y*) and a set of explanatory variables (*x*) which may be continuous or categorical. In this and the next few modules, we consider regression models for *categorical* response variables.

We will consider models for two types of categorical variable (see C1.3.8 for a classification scheme for variables):

- *Nominal*, where the numeric codes assigned to categories are simply labels (e.g. sex, ethnicity)

- *Ordinal*, where the numeric codes imply some ordering (e.g. strength of agreement with a statement in a questionnaire with categories ranging from ‘strongly agree’ to ‘strongly disagree’)

In many subject areas, but especially in the social sciences, categorical responses are more common than continuous responses. In this module, we consider models for a particular type of categorical response – *binary or dichotomous responses*, that is variables with only two categories. Examples include:

- Voting intentions in two-party systems, e.g. Republican vs. Democrat in the US

- Exam performance where only a pass or fail is recorded, e.g. in a driving test

- Mortality or presence of a medical condition

Note that when there are only two categories, it does not matter whether one category can be thought of as ‘higher’ than the other; the distinction between nominal and ordinal is irrelevant. In later modules, we will see how the methods described here can be extended to handle categorical responses with more than two categories. In that case, the distinction between nominal and ordinal is important and we will need to consider different (but closely related) models for each.

To introduce ideas, we will assume in this module that our data do not come from a hierarchically-structured population. However, all methods we describe can be extended to allow for and to explore clustered data and, in Module 7, we will meet multilevel models for binary response data.

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.

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.

The survey covered 49 states, but we restrict our analysis to only three – California, New York and Texas. The total sample size in the selected states is 3688. (In C6.8 and Module 7 we extend the analysis to all 49 states.)

In this module, we consider three explanatory variables:

- Age in years
- Sex (coded 0 for male and 1 for female)
- State (coded 1 for California, 2 for New York and 3 for Texas)

In C6.8 (where we analyse the proportion of respondents who would vote Bush in a state) we consider two explanatory variables:

- Proportion of non-white respondents in the state
- Proportion of respondents who attend religious services at least once a week

C6.1 Preliminaries: Mean and Variance of Binary Data

Denote by y_i the binary response for individual *i*, coded 0 or 1.

Mean of binary y

Recall that the population mean, or *expected value*, of a variable *y* is given by

$$\mu = E(y) = \frac{1}{N}\sum_{i=1}^{N} y_i$$

where *N* is the population size and $\{y_i\}$ are the values of *y* for members of the population.

Suppose that in the population there are *R* individuals with a *y*-value of 1, and therefore *N-R* individuals with a *y*-value of 0. Then the expression for the population mean simplifies to the proportion of individuals with a *y*-value of 1, which we will denote by π:

$$\mu = \pi = \Pr(y = 1) = \frac{R}{N}. \qquad (6.1)$$

When *y* is taken as the response variable in an analysis, we will refer to π as the *response probability*.[2] Some authors refer to π as the *success probability*, where obtaining a *y*-value of 1 is regarded a success and a value of 0 a failure.

Of course, we will not generally know the population mean and we will estimate it by the proportion of individuals with a *y*-value of 1 in our sample:

$$\hat{\pi} = \bar{y} = \frac{r}{n}$$

where *r* and *n* are the sample values of *R* and *N*.

Variance of binary y

Recall also that the population variance (the square of the standard deviation) of a variable *y* is given by

$$\sigma^2 = \text{var}(y) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2 .$$

[2] This should not be confused with the probability of responding in the survey. Here, we use the term response probability for the probability of being in a particular response category (*y* = 1).

For binary *y* we substitute π for μ and, using the facts that i) $y_i^2 = y_i$ (because 1^2 = 1 and 0^2 = 0) which implies $\sum_{i=1}^{N} y_i^2 = \sum_{i=1}^{N} y_i$ and ii) $\sum_{i=1}^{N} y_i = R = N\pi$, we obtain:

$$\sum_{i=1}^{N}(y_i - \pi)^2 = \sum_{i=1}^{N}(y_i^2 - 2\pi y_i + \pi^2) = N\pi - 2N\pi^2 + N\pi^2 = N\pi(1-\pi) .$$

Therefore the variance simplifies to

$$\sigma^2 = \text{var}(y) = \pi(1-\pi). \qquad (6.2)$$

The sample estimate of the variance, denoted by $\hat{\sigma}^2$ or s^2, is

$$s^2 = \hat{\pi}(1-\hat{\pi}) .$$

The Bernoulli and binomial distributions

From (6.1) and (6.2) we can see that the mean and variance for a binary variable *y* are defined by a single parameter π, unlike a continuous *y* which needs two separate parameters to define its mean and variance. A distribution with mean π and variance π(1- π) is called a *Bernoulli distribution*.

Sometimes *y* is said to follow a *binomial* distribution but, strictly, the binomial distribution has an extra parameter that is redundant for binary data. The more general binomial distribution applies to grouped binary data, where instead of observing a binary *y* for each individual we observe the proportion of individuals in a group with the value *y* = 1. In the case of grouped data, we need π and the total number in a group (the *denominator* for the response probability) to define the distribution of the proportion. The Bernoulli distribution is a special case of the binomial distribution with the additional 'denominator' parameter set to 1. (Grouped binary data are the subject of C6.8 at the end of this module.)

Expected value for an individual: towards modelling

For a given individual *i*, their expected value for *y* is denoted by

$$E(y_i) = \pi_i = \Pr(y_i = 1). \qquad (6.3)$$

In the absence of other information $\pi_i = \pi$, i.e. their expected value is simply the response probability for the population (estimated by the sample response probability $\hat{\pi}_i$). More generally, however, an individual's response will depend on their values on a set of explanatory variables x_1, x_2, …, x_p and therefore the expected response will vary across individuals (hence the *i* subscript on π). Our objective in this module is to specify a suitable model that relates an individual's response probability π_i to their values on the *x*s: x_{1i}, x_{2i}, …, x_{pi}.

Don’t forget to do the practical for this section! (see beginning of document for details of how to find the practical)

Please read P6.1, which is available in online form or as part of a pdf file.

Don’t forget to take the online quiz for this section! (see page 2 for details of how to find the quiz questions)

C6.2 Moving towards a Regression Model for *y*: The Linear Probability Model

You may be thinking: why can’t we just use multiple linear regression to analyse binary response data? While it is possible to use multiple regression – called a linear probability model when the response is binary – there are a number of problems with this approach, so it is not generally recommended. The aim of this section is to describe these problems, thereby motivating the need for special techniques.

C6.2.1 Revision of linear regression

Consider the linear regression model for continuous *y* and a single explanatory variable *x* (Module 3):

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad (6.4)$$

where e_i is a residual term representing unobserved characteristics of individual *i* that determine the response y_i after controlling for the effect of x_i.

Because e_i is unobserved we can fix its mean at whatever we like, and we usually assume a zero mean so that β_0 is then the mean of *y* when *x* = 0. Under this assumption, the mean or expected value of an individual’s *y*-value is

$$E(y_i) = \beta_0 + \beta_1 x_i\,. \qquad (6.5)$$

Note that the expected value on the left hand side of (6.5) might more accurately be written $E(y_i|x_i)$, that is “the expected value of y_i given or conditional on x_i”. This notation is used to emphasise the point that we are modelling the expected value of y_i as a function of x_i. In this module we will use the simpler form $E(y_i)$, but bear in mind that in regression we are always modelling the *conditional* mean of y_i (conditional on whatever explanatory variables are included in the model).

So far we have assumed that e_i has zero mean. Two other standard regression assumptions (see C3.1.2) are that the e_i have constant variance σ_e^2, and that they follow a normal distribution. Putting together these assumptions about e_i we have: $e_i \sim N(0, \sigma_e^2)$. Note also that we assume the e_i for different individuals are independent.