Module 6: Regression Models for Binary Responses

R Practical

Camille Szmaragd and George Leckie[1]
Centre for Multilevel Modelling

Pre-requisites box

- Modules 1-3

Contents

[1] This R practical is adapted from the corresponding MLwiN practical: Steele, F. (2008) Module 6: Regression Models for Binary Responses. LEMMA VLE, Centre for Multilevel Modelling. Accessed at http://www.cmm.bris.ac.uk/lemma/course/view.php?id=13.

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

EXAMPLE

From within the LEMMA learning environment

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

Introduction to the Bangladesh Demographic and Health Survey 2004 Dataset

You will be analysing data from the Bangladesh Demographic and Health Survey (BDHS), a nationally representative cross-sectional survey of women of reproductive age (13-49 years).[2]

Our response variable is a binary indicator of whether a woman received antenatal care from a medically-trained provider (a doctor, nurse or midwife) at least once before her most recent live birth. To minimise recall errors, the question was asked only about children born within five years of the survey. For this reason, our analysis sample is restricted to women who had a live birth in the five-year period before the survey. Note that if a woman had more than one live birth during the reference period, we consider only the most recent.

We consider a range of predictors, including the woman's age at the time of the birth, her level of education, and an indicator of whether she was living in an urban or rural area at the time of the survey. The dataset contains the following variables:

Variable name	Description and codes
comm	Community identifier (not used until P6.8)
womid	Woman identifier
antemed	Received antenatal care at least once from a medically-trained provider, e.g. doctor, nurse or midwife (1 = yes, 0 = no)
bord	Birth order of child (ranges from 1 to 13)
mage	Mother's age at the child's birth (in years)
urban	Type of region of residence at survey (1 = urban, 0 = rural)
meduc	Mother's level of education at survey (1 = none, 2 = primary, 3 = secondary or higher)
islam	Mother's religion (1 = Islam, 0 = other)
wealth	Household wealth index in quintiles (1 = poorest to 5 = richest)

[2] We thank MEASURE DHS for their permission to make these data available for training purposes. Additional information about the 2004 BDHS and other Demographic and Health Surveys, including details of how to register for a DHS Download Account, is available from www.measuredhs.com.

P6.1 Preliminaries: Mean and Variance of Binary Data

Download the R dataset for this lesson:

From within the LEMMA Learning Environment
- Go to **Module 6: Regression Models for Binary Responses**, and scroll down to **R Datasets and R files**
- Right click "6.1.txt" and select **Save Link As…** to save the dataset to your computer.

Read the dataset into R using the `read.table` command and create a dataframe object called **mydata**[3]:

```
> mydata <- read.table("6.1.txt", sep = ",", header = TRUE)
```

and use the `str` command to produce a summary of the dataset:

```
> str(mydata)
'data.frame':   5366 obs. of  9 variables:
 $ comm   : int  1 1 1 1 1 1 1 1 1 1 ...
 $ womid  : int  1 2 3 4 5 6 7 8 9 10 ...
 $ antemed: int  0 1 1 0 0 1 0 0 0 1 ...
 $ bord   : int  4 2 3 6 6 4 2 3 1 1 ...
 $ mage   : int  33 21 26 28 37 29 20 29 19 19 ...
 $ urban  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ meduc  : int  2 3 2 1 2 2 3 3 3 3 ...
 $ islam  : int  1 1 1 1 1 1 1 1 1 1 ...
 $ wealth : int  3 4 2 2 4 4 2 3 3 4 ...
```

There are 5,366 women in the dataset.

[3] At the beginning of your R session, you will need to set R's working directory to the file location where you saved the dataset. This can be done using the command line and the `setwd` function:

```
> setwd("C:/userdirectory")
```

Or through selecting Change dir… on the File menu.

P6.1.1 Mean and standard deviation of the response variable

We will begin by tabulating our response variable, **antemed** to obtain the frequencies, percentages and cumulative percentages for **antemed**:

```
> cbind(Freq = table(mydata$antemed), Perc = prop.table(table(mydata$antemed)),
Cum = cumsum(prop.table(table(mydata$antemed))))

  Freq      Perc       Cum
0 2613 0.4869549 0.4869549
1 2753 0.5130451 1.0000000
```

The sample estimate of the proportion of women receiving antenatal care is $\hat{\pi} = 0.513$.[4]

Next, we will calculate a range of summary statistics for **antemed** and also its standard deviation.

```
> summary(mydata$antemed)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   1.000   0.513   1.000   1.000

> sd(mydata$antemed)
[1] 0.4998764
```

Notice that the mean of 0.513 is equal to the proportion receiving antenatal care that we obtained from the tabulation.

Using the formula for the standard deviation of a binary variable given in C6.1, we obtain

$s = \sqrt{\hat{\pi}(1-\hat{\pi})} = \sqrt{0.513(1-0.513)} = 0.4998$, which agrees with the sd value in the output.

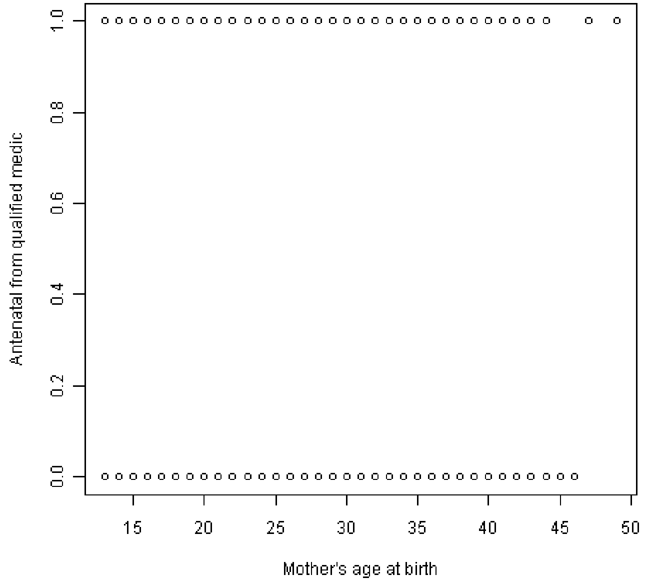
[4] Throughout the practical we will frequently refer to antenatal care from a medically-trained provider simply as antenatal care.

P6.1.2 Bivariate relationships between the response and explanatory variables

Before fitting any models to the relationship between **antemed** and explanatory variables, we will first examine the bivariate relationship between **antemed** and three potential predictors: maternal age (**mage**), type of region of residence (**urban**) and maternal education (**meduc**).

We begin with **mage**, a continuous variable. Let’s start with a scatterplot of **antemed** versus **mage**.

```
> plot(mydata$mage, mydata$antemed, xlab = ”Mother’s age at birth”, ylab =
”Antenatal from qualified medic”)
```



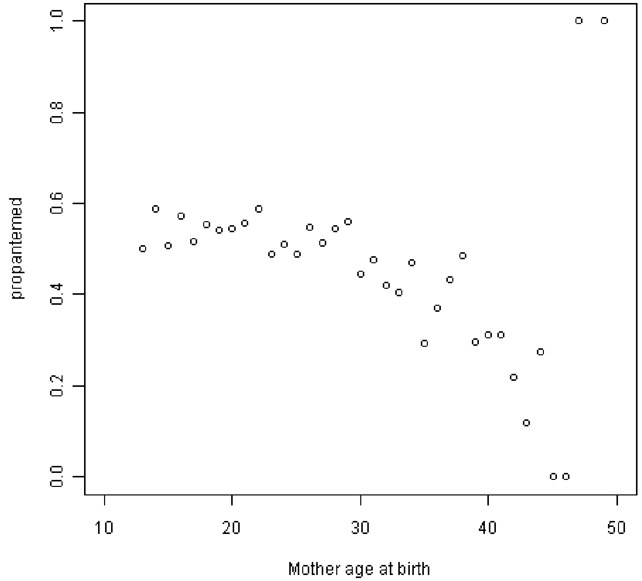
Clearly the scatterplot is not very informative because our response takes only two values. Instead we will plot the proportion receiving antenatal care (i.e. the mean of **antemed**) against **mage**. To do this, we calculate the mean of **antemed** for each distinct value of **mage**. To create a new variable equal to the mean of another variable, we can use the `tapply` command with the `mean` option:

```
> propantemed <- tapply(mydata$antemed, mydata$mage, mean)
```

Here, `tapply` calculates the mean of **antemed** (the first variable in `tapply`) for each different value of **mage**.

We can now repeat the above `plot` command but swap **antemed** for **propantemed**:

```
> plot(names(propantemed), propantemed, xlim = c(10, 50), xlab = "Mother age at
birth")
```



The relationship between the proportion receiving antenatal care and maternal age is fairly linear, but with some curvature at older ages and outliers at the top right of the plot. If you look at the data (**propantemed)** you will see that the outlying points represent only two women who gave birth at ages 47 and 49. We will consider a quadratic function for **mage** in our regression models.

The other two predictors we will consider (**urban** and **meduc**) are categorical, so we will examine their relationship with **antemed** using crosstabulations.

To tabulate **antemed** versus **urban**:

```
> table(mydata$antemed, mydata$urban)

     0    1
  0 2138  475
  1 1544 1209

> table(mydata$antemed, mydata$urban) / (rbind(colSums(table(mydata$antemed,
mydata$urban)), colSums(table(mydata$antemed, mydata$urban))))

          0         1
  0 0.5806627 0.2820665
  1 0.4193373 0.7179335
```

We find that a mother is far more likely to receive antenatal care from a medically-trained provider if she lives in an urban area rather than a rural area (72% compared with 42%).

To tabulate **antemed** versus **meduc**:

```
> table(mydata$antemed, mydata$meduc)

     1    2    3
  0 1272  856  485
  1  594  793 1366

> table(mydata$antemed, mydata$meduc) / rbind(colSums(table(mydata$antemed,
mydata$meduc)), colSums(table(mydata$antemed, mydata$meduc)))

          1         2         3
  0 0.6816720 0.5191025 0.2620205
  1 0.3183280 0.4808975 0.7379795
```

We also find a strong relationship between antenatal care and maternal education; the probability of receiving antenatal care increases from 32% for a mother with no schooling to 74% if she was educated to at least secondary level.

To summarise, there are strong bivariate relationships between antenatal care and all three predictors considered. In the following exercises we will consider the effects of these and other variables jointly in regression analysis.

Don’t forget to take the online quiz!

From within the LEMMA learning environment

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

P6.2 Moving Towards a Regression Model: The Linear Probability Model

Download the R dataset for this lesson:

From within the LEMMA Learning Environment

- Go to **Module 6: Regression Models for Binary Responses**, and scroll down to **R Datasets and R files**
- Right click “6.2.txt” and select **Save Link As…** to save the dataset to your computer

Read the dataset into R:

```
> mydata <- read.table("6.2.txt", sep = ",", header = TRUE)
```

Consider a linear probability model for the probability π_i that mother i receives antenatal care from a medically trained provider, with **mage**, **urban** and **meduc** as explanatory variables. The model has the following form:

$$\textbf{antemed}_i = \pi_i + e_i$$

$$\pi_i = \beta_0 + \beta_1\textbf{mage}_i + \beta_2\textbf{urban}_i + \beta_3\textbf{meduc2}_i + \beta_4\textbf{meduc3}_i$$

where $e_i \sim N(0, \sigma_e^2)$.

meduc2 and **meduc3** are dummy variables for categories 2 and 3 (primary and secondary or higher) of **meduc**, taking category 1 (no education) as the reference. Before we fit this model, we create the dummy variables[5]:

```
> mydata$meduc2 <- mydata$meduc == 2

> mydata$meduc3 <- mydata$meduc == 3
```

We also choose to centre **mage** so that the model intercept can be interpreted as the probability of receiving antenatal care for a mother of mean age. This is done by using the `summary` command to find the grand mean of **mage** and then defining a new variable **magec** that is equal to **mage** minus its grand mean.

```
> summary(mydata$mage)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13.00   19.00   23.00   23.63   28.00   49.00

> mydata$magec <- mydata$mage - mean(mydata$mage)
```

We also define a squared age variable **magecsq** and add this to the model in order to fit a quadratic function for **magec**:

[5] An alternative approach is to convert **meduc** into a `factor` (see module 3 footnote 9 for details).