Module 6: Regression Models for Binary Responses MLwiN Practicals

Fiona Steele
Centre for Multilevel Modelling

Pre-requisites

- Modules 1-3

Contents

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 Q1 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).[1a]

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 file 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)

[1a] 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

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)
cons	A column of ones. This variable will be included as an explanatory variable in all models and its coefficient will be the intercept

There are 5366 women in the data file.

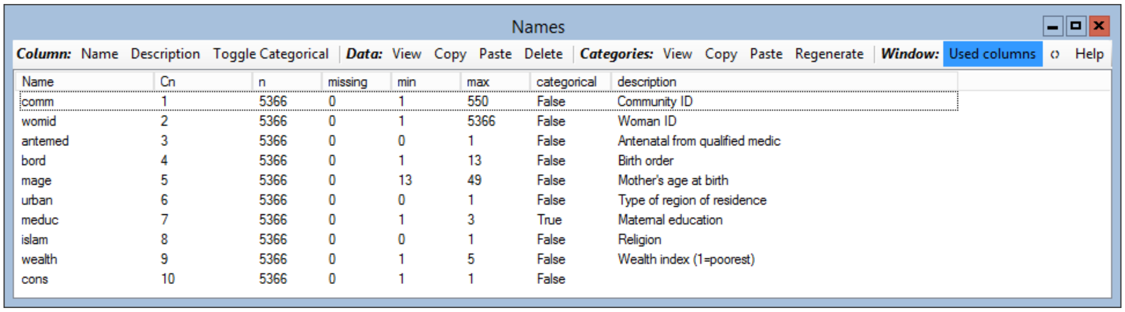
To open the worksheet:

From within the LEMMA Learning Environment

- Go to **Module 6: Regression Models for Binary Responses**, and scroll down to ***MLwiN* Datafiles**
- If you do not already have MLwiN to open the datafile with, click **(get *MLwiN*)**.
- Click “ 6.1.wsz”

The **Names** window will appear.

- Click the check box next to **Used columns** to view only those columns that contain data


Names
Column: Name Description Toggle Categorical
Data: View Copy Paste Delete
Categories: View Copy Paste Regenerate
Window: Used columns
Help

Name	Cn	n	missing	min	max	categorical	description
comm	1	5366	0	1	550	False	Community ID
womid	2	5366	0	1	5366	False	Woman ID
antemed	3	5366	0	0	1	False	Antenatal from qualified medic
bord	4	5366	0	1	13	False	Birth order
mage	5	5366	0	13	49	False	Mother's age at birth
urban	6	5366	0	0	1	False	Type of region of residence
meduc	7	5366	0	1	3	True	Maternal education
islam	8	5366	0	0	1	False	Religion
wealth	9	5366	0	1	5	False	Wealth index (1=poorest)
cons	10	5366	0	1	1	False	

P6.1 Preliminaries: Mean and Variance of Binary Data

P6.1.1 Mean and standard deviation of the response variable

We will begin by tabulating our response variable, **antemed**.

- From the **Basic Statistics** menu, select **Tabulate**
- Check **Percentages of row totals**
- From the drop-down list next to **Columns**, select **antemed**
- Click **Tabulate**

The following table will appear in the Output window:

```
          0         1         TOTALS

     N    2613      2753      5366
     %    48.7      51.3      100.0
```

The sample estimate of the proportion of women receiving antenatal care[1b] is $\hat{\pi}$ = 0.513.

Next, we will calculate the mean and standard deviation of **antemed**.

- From the **Basic Statistics** menu, select **Averages and Correlations**
- Select **antemed** from the variable list
- Click **Calculate**

```
               N         Missing   Mean       s.d.
antemed        5366      0         0.51305    0.49988
```

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 s.d. value in the output.

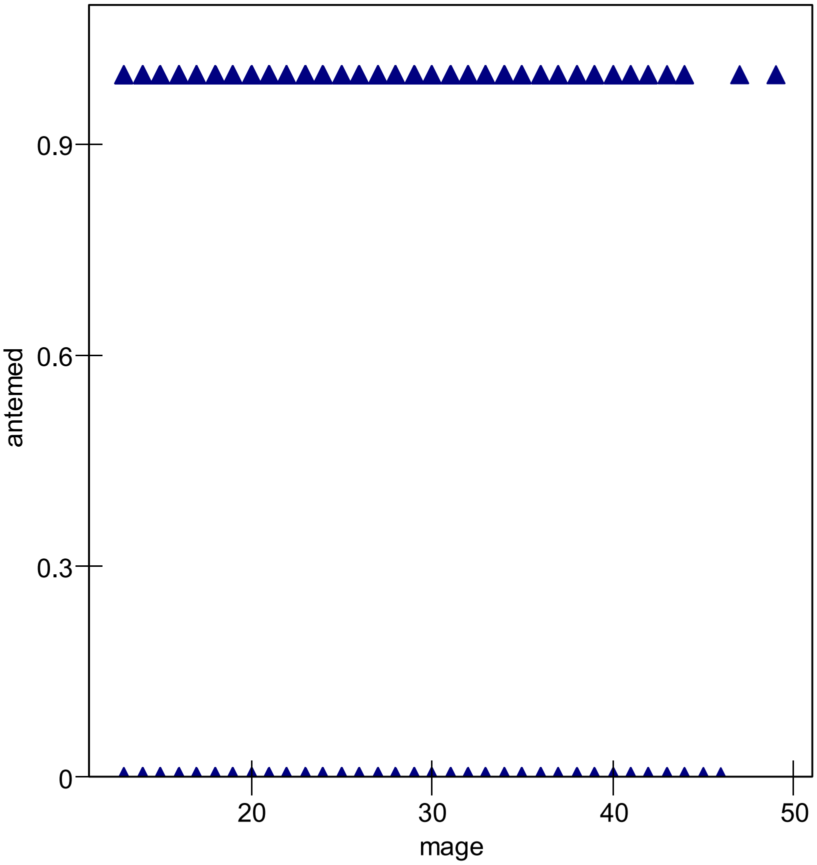
[1b] Throughout the practical we will frequently refer to antenatal care from a medically-trained provider simply as antenatal care.

P6.1.2 Bivariate relationships between the response and explanatory variables

Before fitting any models to the relationship between *y* (**antemed**) and explanatory variables, we will first examine the bivariate relationship between *y* 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**.

- From the **Graphs** menu, select **Customised Graph(s)**
- From the drop-down list labelled **plot type**, select **point**
- From the drop-down list labelled **y**, select **antemed**
- From the drop-down list labelled **x**, select **mage**
- Click **Apply**
- You can add titles by clicking anywhere on the graph and selecting the **Titles** tab



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**, but first we need to sort the values of **antemed** by **mage**. We will store the sorted values of **antemed** and **mage** in columns c11 and c12, which we will call **ante-sort** and **mage-sort**.

- From the **Data Manipulation** menu, select **Sort**
- Under **Key code columns**, select **mage**
- Under **Input columns**, highlight **antemed** and **mage** (using Ctrl-click)
- Under **Output columns**, click **Free columns** (so that the next empty columns, c11 and c12 will be used)
- Click **Add to action list** followed by **Execute**
- Go to the **Names** window and type in the names **ante-sort** and **mage-sort** for the new variables in **c11** and **c12**

We can now calculate the mean of **ante-sort** for each value of **mage-sort**.

- From the **Data Manipulation** menu, select **Multilevel data manipulations**
- Under **Operation**, retain the default of **Average**
- Under **On blocks defined by**, select **mage-sort**
- Under **Input columns**, highlight **ante-sort** and **mage-sort**
- Under **Output columns,** click **Same as input**
- Click **Add to action list** followed by **Execute**

If you look at **ante-sort** and **mage-sort**(using **Data Manipulation** → **View or edit data**) you will see that values of **ante-sort** are the means for each value of **mage-sort**.

We are now in a position to plot the mean of **antemed** (proportion receiving antenatal care) versus maternal age.

- From the **Graphs** menu, select **Customised Graph(s)**
- From the drop-down list labelled **plot type**, select **point**
- From the drop-down list labelled **y**, select **ante-sort**
- From the drop-down list labelled **x**, select **mage-sort**
- Click **Apply**
- Click anywhere on the plot and then on the **Titles** tab. Change the **y title** to **mean(antemed)**.