Module 6: Regression Models for Binary Responses

Stata Practical

George Leckie [1]
Centre for Multilevel Modelling

Pre-requisites box

- Modules 1-3

Contents

[1] This Stata 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

Load "6.1.dta" into memory and open the do-file for this lesson:

From within the LEMMA Learning Environment
- Go to **Module 6: Regression Models for Binary Responses**, and scroll down to ***Stata* Datasets and Do-files**
- Click "6.1.dta" to open the dataset

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

```
. describe

Contains data from 6.1.dta
  obs:         5,366
 vars:             9                          5 Sep 2009 09:38
 size:       101,954 (99.9% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
comm            int    %9.0g                  Community ID
womid           int    %9.0g                  Woman ID
antemed         byte   %9.0g                  Antenatal from qualified medic
bord            byte   %9.0g                  Birth order
mage            byte   %9.0g                  Mother's age at birth
urban           byte   %9.0g                  Type of region of residence
meduc           byte   %9.0g                  Maternal education
islam           byte   %9.0g                  Religion
wealth          byte   %9.0g                  Wealth index (1 = poorest)
-------------------------------------------------------------------------------
Sorted by:
```

There are 5,366 women in the dataset.

P6.1.1 Mean and standard deviation of the response variable

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

```
. tabulate antemed

  Antenatal |
       from |
  qualified |
      medic |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      2,613       48.70       48.70
          1 |      2,753       51.30      100.00
------------+-----------------------------------
      Total |      5,366      100.00
```

The sample estimate of the proportion of women receiving antenatal care is $\hat{\pi} = 0.513$.[3]

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

```
. summarize antemed

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     antemed |      5366    .5130451    .4998764          0          1
```

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 `Std. Dev.` value in the output.

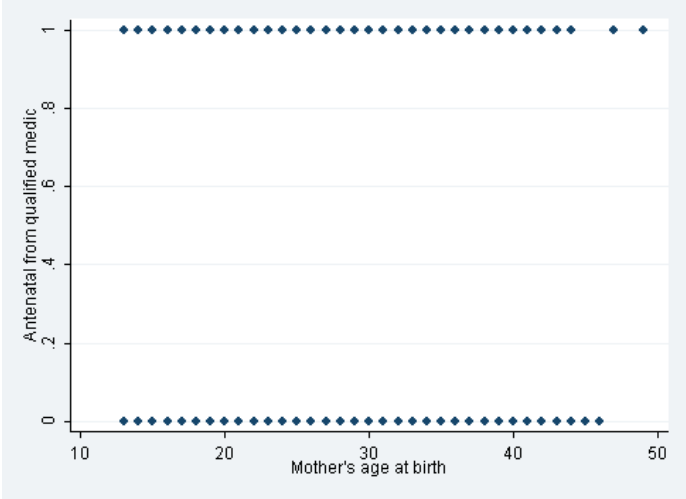
[3] Throughout the practical we will frequently refer to antenatal care from a medically-trained provider simply as antenatal care.

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**.

```
. scatter antemed mage
```

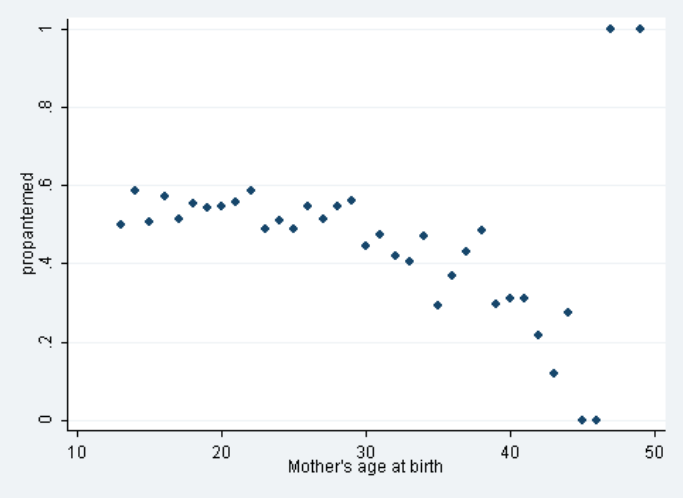


Clearly the scatterplot is not very informative because our response takes only two values. Instead we will plot the proportion receiving antenatal care (i.e. the mean of **antemed**) against **mage**. To do this, we calculate the mean of **antemed** for each distinct value of **mage**. To create a new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

```
. bysort mage: egen propantemed = mean(antemed)
```

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

```
. scatter propantemed mage
```



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 (`list mage propantemed, sepby(mage)`) 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**:

```
. tabulate antemed urban, column

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| column percentage  |
+--------------------+

 Antenatal |
      from |   Type of region of
 qualified |        residence
     medic |         0          1 |     Total
-----------+----------------------+----------
         0 |     2,138        475 |     2,613
           |     58.07      28.21 |     48.70
-----------+----------------------+----------
         1 |     1,544      1,209 |     2,753
           |     41.93      71.79 |     51.30
-----------+----------------------+----------
     Total |     3,682      1,684 |     5,366
           |    100.00     100.00 |    100.00
```

Where we use the `column` option to report column percentages (they sum to 100% going down the table) in addition to the cell frequencies. 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**:

```
. tabulate antemed meduc, column

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| column percentage  |
+--------------------+

 Antenatal |
      from |
 qualified |        Maternal education
     medic |         1          2          3 |     Total
-----------+---------------------------------+----------
         0 |     1,272        856        485 |     2,613
           |     68.17      51.91      26.20 |     48.70
-----------+---------------------------------+----------
         1 |       594        793      1,366 |     2,753
           |     31.83      48.09      73.80 |     51.30
-----------+---------------------------------+----------
     Total |     1,866      1,649      1,851 |     5,366
           |    100.00     100.00     100.00 |    100.00
```

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.

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.