Biostatistics 140.655, 2017-18
Lab 3 Solution

Topics:

- Refresher on logistic regression
- Exploration of correlated binary data

Learning Objectives:

Students who successfully complete this lab will be able to:

- Feel confident interpreting a regression coefficient from a logistic regression model
- Explore the marginal prevalence of binary outcomes over time
- Explore the within subject patterns of binary outcomes over time

Associated Quiz:

- While we will review and discuss parts of this exercise, there is a short quiz (Quiz 3) on Courseplus which will assess your basic knowledge of the course materials thus far with focus on ideas from this lab session.
- Quiz 3 is available on Courseplus Wednesday Feb 21st, please complete the quiz by 5pm Friday Feb 23rd.

Scientific Background:

You are involved with a randomized control trial of a new surgical procedure vs. standard medical care for the treatment of stroke. The goal is to improve functional disability among the stroke patients. The trial will use a scale that rates the patients' functional disability including mortality status. We will consider a binary representation of the scale such that a 1 indicates "little to no disability" and 0 indicates "moderate to severe disability or death."

The functional disability is assessed at baseline (time = 0) and then at 6, 12 and 18 months.

The data consist of:

- A = treatment assignment, 0 = standard medical care, 1 = surgical intervention
- Y = binary indicator for functional disability, 0 = moderate to severe disability or death, 1 = little to no disability
- Time = 0, 6, 12, and 18 months.

<u>Review of logistic regression to be completed prior to the lab session:</u>

Logistic regression can be motivated through the use of 2 x 2 tables.

Suppose the study described above consists of 200 stroke patients randomized 1:1 to the new surgical procedure (A = 1) vs. standard medical care (A = 0). In this section, the outcome we will consider is the measure of functional disability at 18-month follow-up. Let Y = 1 indicate “little to no disability” and Y = 0 indicates “moderate to severe disability or death” measured 18-months following treatment.

The effect of the new surgical procedure may be modified by the location of the stroke. Let W be the binary moderator; W = 1 indicating the stroke occurred in the thalamus vs. W = 0 indicating other locations.

We have simulated data from this trial and presented the results stratified by the potential moderator W (see simulate_binary.do if you are interested in the simulation code).

```
-> W = 0

+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+

           |           Y
         A |         0          1 |     Total
-----------+----------------------+----------
         0 |        33         19 |        52 
           |     63.46      36.54 |    100.00 
-----------+----------------------+----------
         1 |        32         25 |        57 
           |     56.14      43.86 |    100.00 
-----------+----------------------+----------
     Total |        65         44 |       109 
           |     59.63      40.37 |    100.00 

-> W = 1


+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+

           |           Y
         A |         0          1 |     Total
-----------+----------------------+----------
         0 |        29         19 |        48 
           |     60.42      39.58 |    100.00 
-----------+----------------------+----------
         1 |        17         26 |        43 
           |     39.53      60.47 |    100.00 
-----------+----------------------+----------
     Total |        46         45 |        91 
           |     50.55      49.45 |    100.00 
```

The pre-specified analysis plan for the trial is as follows:

1. Estimate the OR for "little to no disability" comparing the new surgical procedure to standard medical care among patients whose stroke was located outside the thalamus (W = 0)

2. Determine if the strength of the association between functional disability and treatment assignment differs across the patients whose stroke was located in the thalamus vs. outside of the thalamus.

To conduct this pre-specified analysis plan, you fit the following logistic regression model:

$$Log\left(\frac{\Pr(Y_i=1|A_i,W_i)}{\Pr(Y_i=0|A_i,W_i)}\right) = \beta_0 + \beta_1 A_i + \beta_2 W_i + \beta_3 A_i W_i, \text{ where } i \text{ denotes subject i}$$

Do the following:

A. Interpret each of the coefficients in the model above.

β_0: Log odds for "little to no disability" among patients whose stroke was located outside the thalamus and who received standard medical care.

β_1: Log odds ratio for "little to no disability" among patients whose stroke was located outside the thalamus comparing those who received the new surgical procedure to those who received standard medical care.

β_2: Log odds ratio for "little to no disability" among patients who received standard medical care comparing those with stroke in the thalamus to those with stoke outside the thalamus.

β_3: Difference in log odds ratio for "little to no disability" given by (log odds ratio among patients whose stroke was in the thalamus comparing those who received the new surgical procedure to those who received standard medical care) minus (log odds ratio among patients whose stroke was outside the thalamus comparing those who received the new surgical procedure to those who received standard medical care).

B. Using the data from the two 2x2 tables presented above, estimate β_0, β_1, β_2, and β_3.

$$\hat{\beta}_0 = \log\left(\frac{19}{33}\right) = -0.552$$

$$\hat{\beta}_1 = \log\left(\frac{25}{32}\right) - \log\left(\frac{19}{33}\right) = \log\left(\frac{25/32}{19/33}\right) = 0.305$$

$$\hat{\beta}_2 = \log\left(\frac{19}{29}\right) - \log\left(\frac{19}{33}\right) = \log\left(\frac{19/29}{19/33}\right) = 0.129$$

$$\hat{\beta}_3 = \log\left(\frac{26/17}{19/29}\right) - \log\left(\frac{25/32}{19/33}\right) = \log\left(\frac{\frac{26/17}{19/29}}{\frac{25/32}{19/33}}\right) = 0.543$$

C. Confirm that your estimates computed by hand, match those obtained via the logistic regression.

```
. logit Y A W A##W

Logistic regression                               Number of obs   =        200
                                                  LR chi2(3)      =       6.25
                                                  Prob > chi2     =     0.1001
Log likelihood = -134.29226                       Pseudo R2       =     0.0227

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           A |   .3052085   .3926635     0.78   0.437    -.4643978    1.074815
           W |   .1292117   .4123699     0.31   0.754    -.6790185    .9374419
         1.A |          0  (omitted)
         1.W |          0  (omitted)
             |
         A#W |
        1 1  |   .5425315   .5818797     0.93   0.351    -.5979318    1.682995
             |
       _cons |  -.5520686   .2879837    -1.92   0.055    -1.116506    .0123691
------------------------------------------------------------------------------

.  logistic Y A W A##W

Logistic regression                               Number of obs   =        200
                                                  LR chi2(3)      =       6.25
                                                  Prob > chi2     =     0.1001
Log likelihood = -134.29226                       Pseudo R2       =     0.0227

------------------------------------------------------------------------------
           Y | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           A |   1.356908   .5328082     0.78   0.437     .6285135     2.92945
           W |   1.137931   .4692485     0.31   0.754     .5071145    2.553441
         1.A |          1  (omitted)
         1.W |          1  (omitted)
             |
         A#W |
        1 1  |   1.720357   1.001041     0.93   0.351     .5499479    5.381649
             |
       _cons |   .5757576   .1658088    -1.92   0.055     .3274217    1.012446
------------------------------------------------------------------------------
```

Lab Exercises: Exploration of correlated binary data

Now we will be utilizing data from all the follow-ups (times 0, 6, 12 and 18 months).

The objective of the trial is to determine if the prevalence of "little to no disability" over time differs across the treatment groups; with larger improvements over time in functional disability for patients receiving the new surgical procedure.

We have simulated data to represent the clinical trial described above.

We have provided you with the code to simulate correlated binary data using both Stata and R (see simulate_binary.do and simulate_binary.R). In the lab, you will be analyzing the data generated from the Stata code. See stroke_trial.dta (version 13), stroke_trial_old.dta (version 10-12) or stroke_trial.csv.

We have provided you with the STATA and R commands posted below as well as some additional code in Lab3.R and Lab3.do.

1. Create a tabular and graphical display of the prevalence of "little to no disability" at each follow-up for each treatment group.

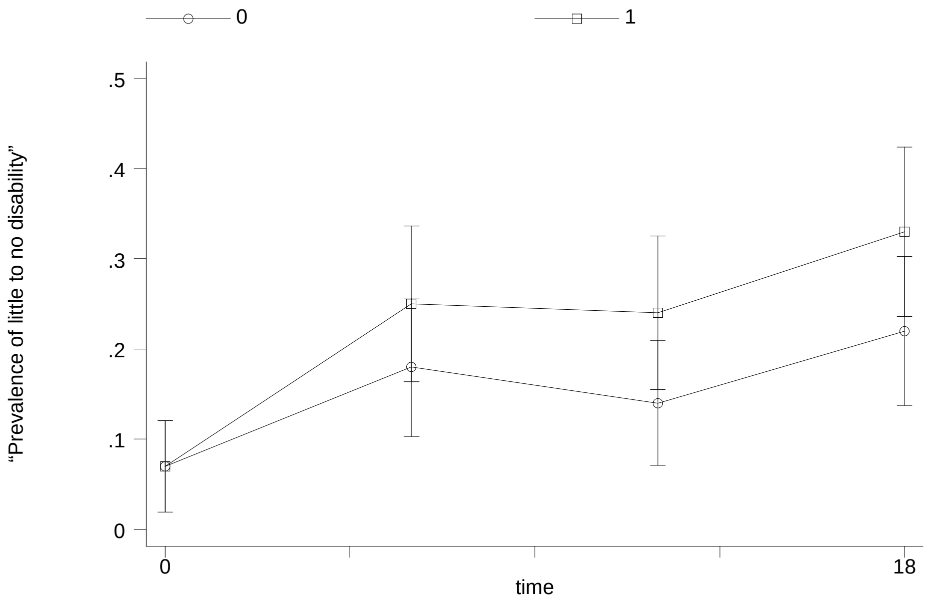
   STATA:

   ```
   tabulate A time, summarize(y) means

   set scheme s1mono
   xtset id time
   xtgraph y, group(A) ylab(0(0.1)0.5) l1("Prevalence of little to no
   disability")
   ```

   R:

   ```
   tapply(data$y,list(data$A,data$time),mean)

   library(ggplot2)
   ggplot(data, aes(x=time, y=y, colour=A)) +
     geom_errorbar(aes(ymin=y-se, ymax=y+se), width=.1) +
     geom_line() +
     geom_point()
   ```

   Based on your descriptive analysis of the prevalence, do you think the prevalence of "little to no disability" is increasing faster over time among subjects receiving the new surgical procedure compared to those receiving standard medical care?



The prevalence of "little to no disability" is not increasing faster among subjects receiving the new surgical procedure as compare to those receiving the standard medical care. Although the prevalence estimates at each time point after the baseline visit are higher for subjects receiving the new surgical procedure, as compared to subjects receiving the standard of care, the 95% confidence intervals for both groups overlap with one another at each time point.

Assume you will be fitting a logistic regression model to the data, is there another figure you are interested in constructing? Remember the logistic regression is modeling the log odds of P(Y = 1) as a function of treatment (A) and time.

We could make a plot of the log odds of "little to no disability" as a function of time and treatment (A); this would allow us to determine if it was reasonable to assume that the log odds of "little to no disability" changed over time linearly or non-linearly.

Write down a logistic regression model that can test the hypothesis that the change in prevalence of "little to no disability" over time is different across the two treatment groups.

$Log\left(\frac{\Pr(Y_{ij}=1)}{\Pr(Y_{ij}=0)}\right) = \beta_0 + \beta_1 I(time_{ij} = 6) + \beta_2 I(time_{ij} = 12) + \beta_3 I(time_{ij} = 18) + \beta_4 A_i I(time_{ij} = 6) + \beta_5 A_i I(time_{ij} = 12) + \beta_6 A_i I(time_{ij} = 18)$, where i denotes subject i and j denotes visit, A_i is the binary indicator for treatment vs. standard medical care, $I(expression)$ is 1 when expression evaluates to TRUE and 0 otherwise.

2. Next, we will start to explore the within subject correlation structure using some commands available within Stata. These include "xtrans" and "xttab."

```
. xttrans y

           |          y
         y |         0          1 |     Total
-----------+----------------------+----------
         0 |     81.98      18.02 |    100.00
         1 |     52.63      47.37 |    100.00
-----------+----------------------+----------
     Total |     77.33      22.67 |    100.00
```

In the output from the "xttrans" command, the rows represent the outcome at time *t* and the columns represent the outcome at time *t+1*. The data are pooled across all the possible transitions from *t* to *t+1* (0 vs. 6 months, 6 vs. 12 months, 12 vs. 18 months). The values in the table are the conditional probability of being in either column given the data came from row 1 or 2.

How often is the transition from "moderate to severe disability or death" to "little to no disability" observed?

"Moderate to severe disability or death" refers to Y=0. "Little to no disability" refers to Y=1. We first start with the rows and look for Y=0. Then, we look at the columns and find Y=1. This is 18.02.

Among follow-up visits where the patient has "moderate to severe disability or death", 18 percent of those follow-up visits are followed by the patient reporting "little to no disability".

How likely is a patient to remain with "little to no disability" across two follow-ups?

"Little to no disability" refers to Y=1. The transition state is from "little to no disability" to "little to no disability", so this is Y=1 for the rows and Y=1 for the columns.

Among follow-ups where patients report "little to no disability", the patient continues to report "little to no disability" for 47% of the subsequent follow-ups.

```
. xttab y

                  Overall             Between            Within
        y |    Freq.  Percent      Freq.  Percent        Percent
----------+-----------------------------------------------------
        0 |      650     81.25        198     99.00          82.07
        1 |      150     18.75         91     45.50          41.21
----------+-----------------------------------------------------
    Total |      800    100.00        289    144.50          69.20
                               (n = 200)
```

In the "xttab" output, we get several pieces of information.

- First, the overall proportion of follow-ups where patients have "little to no disability." In our example, for roughly 19% of the follow-ups, patients report "little to no disability."

- The “Between” provides the proportion of women who EVER experience “little to no disability.” In our example, roughly 45% of the patients EVER report “little to no disability.”
- The “Within” provides a measure of the frequency of seeing each possible outcome within a subject IF the subject EVER had that outcome. In our example, among the roughly 45% of women who EVER report “little to no disability,” they individually report “little to no disability” for 41% of their follow-ups on average.

From this output, answer the following questions:

What percentage of the patients EVER report “moderate to severe disability or death”?

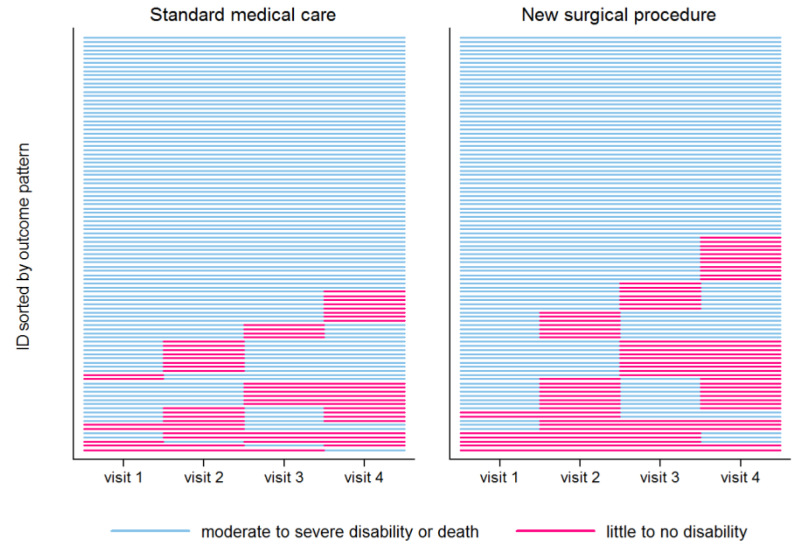
We would look at the “Between” columns to answer this question. “Moderate to severe disability or death” corresponds to Y=0. There were 99.0% of patients who EVER reported “moderate to severe disability or death.”

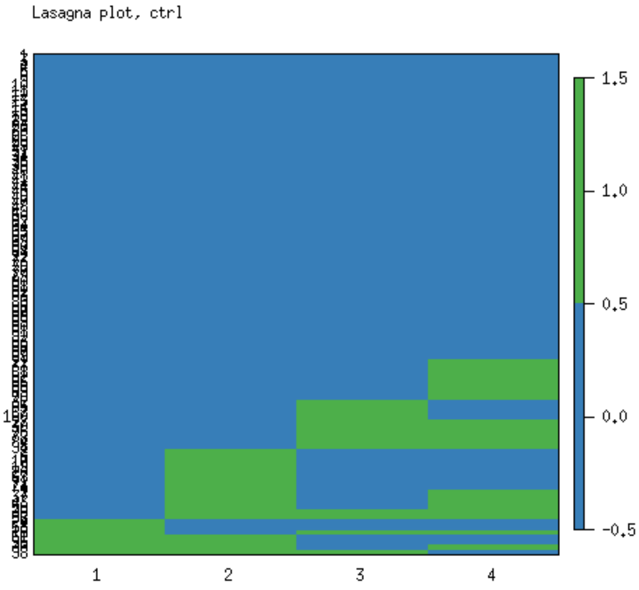
Among patients that EVER report “moderate to severe disability or death”, what is the average percentage of the follow-ups that they report “moderate to severe disability or death”?

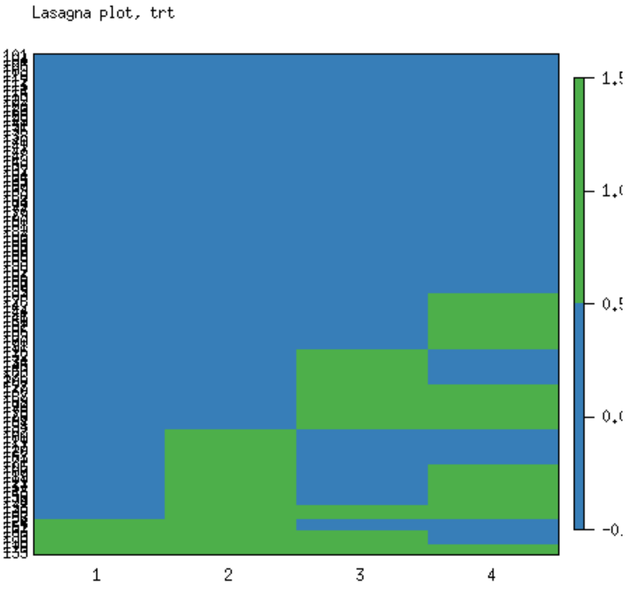
We would look at the “Within” columns to answer this question. “Moderate to severe disability or death” corresponds to Y=0. Among the 99.0% of patients who EVER reported “moderate to severe disability or death,” they individually reported “moderate to severe disability” for 82.1% of their follow-ups on average.

3. The lasagna plot is a “saucy alternative to spaghetti plots” for longitudinal categorical outcomes. The goal is to visualize the trajectories of responses over time but where the responses are categorical. The lasagna plot is constructed with a separate line for each subject; this line is color coded based on the observed response for that subject at each time point. These subject specific lines are then sorted and stacked to allow you to understand common patterns of how the subject’s responses changed over time. With categorical responses, there are a discrete number of potential trajectories.

   We have constructed the lasagna plot for our simulated study below, stratified by treatment group. See the Lab3_2018.do and Lab3_2018.R files for the code to construct these plots.







NOTE: We are "eye-balling" the lasagna plots to answer the questions below.

Roughly what proportion of the patients in the new surgical group remain with "moderate to severe disability or death" for the entire study?

Roughly 50% of the patients receiving the new surgical group remain with "moderate to severe disability or death" for the entire study.

Roughly what proportion of the patients in the standard medical group remain with "moderate to severe disability or death" for the entire study?

Roughly 60% of the patients receiving standard medical care remain with “moderate to severe disability or death” for the entire study.

Roughly what proportion of the patients in the new surgical group move from “moderate to severe disability or death” to “little to no disability” over the course of the study? At what follow-up do most of the transitions occur?

Roughly 45% of the patients in the new surgical group move from “moderate to severe disability or death” to “little to no disability” at least once during the course of the study.

It appears that roughly the same proportion of patients transition at each of the post baseline follow-ups.

Roughly what proportion of the patients in the new surgical group move from “moderate to severe disability or death” to “little to no disability” over the course of the study? At what follow-up do most of the transitions occur?

Roughly 35% of the patients receiving standard medical care move from “moderate to severe disability or death” to “little to no disability” at least once during the course of the study.

It appears that roughly the same proportion of patients transition at each of the post baseline follow-ups.

4. For binary variables, you can compute a correlation coefficient but it is hard to interpret because the value of the correlation coefficient is not bounded between -1 to 1 (we will discuss this further in class). An alternative to computing the correlation coefficient for longitudinal binary data is to compute the pairwise odds ratio as a measure of correlation.

   Reshape your data to wide format and compute all possible pairwise odds ratios (you may want to delegate this task across your group).

   Fill in the table below and comment on any pattern you observe in these measures of correlation.

	Time = 6	Time = 12	Time = 18
Time = 0	17.65	5.00	1.06
Time = 6		1.66	3.03
Time = 12			6.23

Stata:

```
reshape wide y time, i(id) j(followup)
cs y1 y2, or
```

R:

```
data = read.csv("stroke_trial.csv")
data_wide = reshape(data[, colnames(data) != "followup"],timevar = c("time"),
idvar = c("id","A"),direction="wide")

mytable = table(data_wide[,c("y.0","y.6")])
or = mytable[1,1]*mytable[2,2]/(mytable[1,2]*mytable[2,1])
```

5. Similar to marginal models for continuous responses, marginal logistic models require us to specify a model for the mean and a model for the within subject correlation structure. Specify a marginal logistic regression model for this study.

$Y_{ij} \sim Binomial(1, p_{ij}), p_{ij} = \Pr(Y_{ij} = 1 | A_i, time_{ij})$, where *i* denotes subject i and *j* denotes visit

MEAN MODEL:

$$Log\left(\frac{\Pr(Y_{ij} = 1 | A_i, time_{ij})}{\Pr(Y_{ij} = 0 | A_i, time_{ij})}\right)$$
$$= \beta_0 + \beta_1 I(time_{ij} = 6) + \beta_2 I(time_{ij} = 12) + \beta_3 I(time_{ij} = 18) + \beta_4 A_i I(time_{ij} = 6) + \beta_5 A_i I(time_{ij} = 12) + \beta_6 A_i I(time_{ij} = 18)$$

VARIANCE MODEL:

$$Var(Y_{ij}) = p_{ij} \times (1 - p_{ij}), Corr(Y_{ij}, Y_{ik}) = \rho_{|time_{ij} - time_{ik}|}$$