Interpreting categorical predictors with logistic regression using Stata

Joseph M Hilbe 22Jun, 2014
hilbe@asu.edu: © J Hilbe 2014

Modeling categorical predictors with logistic regression can be confusing. I shall briefly describe how to set up and interpret categorical predictors within the context of logistic regression.

Using the Stata **heart** dataset, we shall model *death* (0,1) on four levels of age group. Death is taken to mean death within 48 hours of admission to a hospital for a presumed heart related problem. The age groups were created from an *agegrp* variable distributed as:

```
. tab agegrp, miss

     agegrp |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      1,907       35.39       35.39
          2 |      1,390       25.80       61.19
          3 |      1,425       26.45       87.64
          4 |        666       12.36      100.00
------------+-----------------------------------
      Total |      5,388      100.00
```

The *agegrp* variable is an earlier factoring of the continuous age variable. The indicator (or dummy) variables for each level of *agegrp* may be created using Stata's command

```
. tab agegrp, gen(age)

The following variables (0,1) are created.

age1  =<60      age2  60-69      age3  70-79     age4   >=80
```

We shall use the default level 1 (age1) as the reference. It has the most observations, and represents a reasonable level with which to make comparisons with other higher-aged levels.

```
. glm death age2-age4, fam(bin) nolog

Generalized linear models                          No. of obs      =      5365
Optimization     : ML                              Residual df     =      5361
                                                   Scale parameter =         1
Deviance         =  1805.548188                    (1/df) Deviance =  .3367932
Pearson          =  5364.999729                    (1/df) Pearson  =  1.000746

Variance function: V(u) = u*(1-u)                  [Bernoulli]
Link function    : g(u) = ln(u/(1-u))              [Logit]

                                                   AIC             =  .3380332
Log likelihood   = -902.7740939                    BIC             = -44232.85
------------------------------------------------------------------------------
             |                 OIM
       death |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        age2 |   .7229784   .2659335     2.72   0.007     .2017583    1.244198
        age3 |   1.711221   .2312628     7.40   0.000     1.257954    2.164488
        age4 |   2.454884   .2354266    10.43   0.000     1.993456    2.916311
       _cons |  -4.350278   .2054368   -21.18   0.000    -4.752927   -3.947629
------------------------------------------------------------------------------
```

There is evidence that there is no extra correlation in the data (Pearson.dof=1.00), and it appears that the model is well fitted. We would have to subject it to other goodness of fit tests though (I did outside this note, it appears to be a well fitted logistic model.

The coefficients are interpreted as log-odds; e.g., there is an approximate 2.5 greater log-odds of death among patients aged 80 and over compared to patients under 60 --- among patients in this study. Let us determine the odds ratios of the effects of age groups 2-4 compared to the reference level (age1). The *eform* option exponentiates the coefficients and confidence intervals. The standard errors are calculated using the delta method [OR*se(Beta)].

```
. glm death age2-age4, fam(bin) nolog eform

Generalized linear models                          No. of obs      =       5365
Optimization     : ML                              Residual df     =       5361
                                                   Scale parameter =          1
Deviance         =  1805.548188                    (1/df) Deviance =  .3367932
Pearson          =  5364.999729                    (1/df) Pearson  =  1.000746

Variance function: V(u) = u*(1-u)                  [Bernoulli]
Link function    : g(u) = ln(u/(1-u))              [Logit]
                                                   AIC             =  .3380332
Log likelihood   = -902.7740939                    BIC             = -44232.85
------------------------------------------------------------------------------
             |                 OIM
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        age2 |   2.060561   .5479723     2.72   0.007     1.223552    3.470152
        age3 |   5.535715   1.280205     7.40   0.000     3.518216    8.710137
        age4 |   11.64508   2.741561    10.43   0.000      7.34086    18.47302
       _cons |   .0129032   .0026508   -21.18   0.000     .0086264    .0193004
------------------------------------------------------------------------------
```

We can now state that for patients in this study, without adjustment for other possible confounders, the odds ratio for patients age 60-69 is 2.06. This means that the odds of death within 48 hours of admission to a hospital are some 2 times greater than the odds of death among patients under 60. This is the ratio aspect of the relationship – it is a comparison of the odds of death for one level to the odds or death for the reference level. We may change the reference level if there is a good reason to do so. For instance, to change the reference to age4, simply exclude it from the command line.

We may agegrp to produce the same results as creating indicator variables in memory. Preface *agegrp* with ".i". For instance, I can use the following code with the *nohead* option to retard the display of the header statistics:

```
. glm death i.agegrp, fam(bin) nolog nohead eform
------------------------------------------------------------------------------
             |                 OIM
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agegrp |
          2  |   2.060561   .5479723     2.72   0.007     1.223552    3.470152
          3  |   5.535715   1.280205     7.40   0.000     3.518216    8.710137
          4  |   11.64508   2.741561    10.43   0.000      7.34086    18.47302
             |
       _cons |   .0129032   .0026508   -21.18   0.000     .0086264    .0193004
------------------------------------------------------------------------------
```

The results are identical.

It is easy to change reference levels using this method. For example, if I want to use the highest level as the reference (SAS does this as the default), I will type

```
. glm death ib4.agegrp, fam(bin) nolog eform nohead

------------------------------------------------------------------------------
             |                 OIM
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agegrp |
         1   |   .0858732   .0202168   -10.43   0.000      .054133    .1362238
         2   |    .176947   .0361499    -8.48   0.000     .1185617    .2640839
         3   |   .4753695   .0744064    -4.75   0.000     .3497828    .6460472
             |
       _cons |   .1502591   .0172774   -16.48   0.000     .1199405    .1882415
------------------------------------------------------------------------------
```

The interpretation is a bit more difficult since the odds ratios of lower levels or *agegrp* are smaller than the highest level. Recalling that the odds ratio of *agegrp*=4 was 11.64508 when the reference level was agegrp==1, we can see the relationship it has with *agegrp*==1 when *agegrp*==4 is the reference.

The exponentiated level 1 coefficient may be displayed using the code:

```
. di exp(_b[1.agegrp])
.0858732
```

Now just invert it to see the relationship with the reference level 4.

```
. di 1/exp(_b[1.agegrp])
11.645077
```

In any case we may interpret the above model for *agegrp*==1 as: the odds of death for patients under 60 are some eight-and-a-half times that of the odds of patients aged eighty and greater. Patients aged 70-79 have some half the odds of death within 48 hours of hospital admission compared to patients aged 80 and over.

Remember, when other adjustors are in the model the interpretation of the odds or a level compared to the odds of the reference is taken with the understanding that the other predictor or predictors in the model are held constant.