Perform logistic regression model with interaction between binary variables in Stata

Chaochen Wang | 王　超辰　　Hiroshi Yatsuya | 八谷　寛

2020-10-22 14:46:38 JST created, 2020-10-22 16:13:02 updated

This demonstration can be found with more detailed explanation about how to visually show the interaction effect among categorical variables from the UCLA website:

https://stats.idre.ucla.edu/stata/faq/how-can-i-understand-a-categorical-by-categorical-interaction-in-logistic-regression-stata-12/

You can use the complete command to load the data into your Stata environment:

```
use https://stats.idre.ucla.edu/stat/data/logit2-2, clear
```

The example dataset called logit2-2 includes two binary variables, `f` and `h`, and a continuous variable as covariate `cv1`. We build a model include `f` by `h` interaction, with the covariate `cv1`.

```
##
## . use https://stats.idre.ucla.edu/stat/data/logit(highschool and beyond (200 cases))
##
## . logistic y f##h cv1
##
## Logistic regression                                     Number of obs     =        200
##                                                         LR chi2(4)        =     106.10
##                                                         Prob > chi2       =     0.0000
## Log likelihood =  -78.74193                             Pseudo R2         =     0.4025
##
## ------------------------------------------------------------------------------
##            y | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
## -------------+----------------------------------------------------------------
##          1.f |   20.00771   15.04885     3.98   0.000      4.58104    87.38374
##          1.h |   10.92345   7.218757     3.62   0.000     2.991185     39.8911
##              |
##          f#h |
##         1 1  |   .1290242   .1136444    -2.32   0.020      .022958    .7251177
##              |
##          cv1 |   1.217106   .0399841     5.98   0.000     1.141208    1.298052
##        _cons |   7.06e-06   .0000134    -6.26   0.000     1.72e-07    .0002902
## ------------------------------------------------------------------------------
## Note: _cons estimates baseline odds.
##
## .
```

As you can see all of the variables in the above model including the interaction term are statistically significant. Which means the coefficients in what we fitted in the above model were all statistically significant. The model can be written as below:

$$\text{logit}(Pr(y = 1)) = \alpha + \beta_1 f_i + \beta_2 h_i + \beta_3 f_i \times h_i + \beta_4 \text{cv1}$$

We store the above results in object called `inter`. And build another model without the interaction term (as `main`) and use `lrtest` commend to test the significance of the interaction. Note that the `quietly` is to suppress the output of the `inter` model to save space. For completeness, we will also use a Wald test (`test` command). But we know that a Wald test is an approximation to the likelihood ratio test (`lrtest`), the LRtest is preferred.

```
##
## . use https://stats.idre.ucla.edu/stat/data/logit(highschool and beyond (200 cases))
##
## . logistic y i.f i.h cv1
##
## Logistic regression                               Number of obs     =        200
##                                                   LR chi2(3)        =     100.26
##                                                   Prob > chi2       =     0.0000
## Log likelihood =   -81.6618                       Pseudo R2         =     0.3804
##
## ------------------------------------------------------------------------------
##            y | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
## -------------+----------------------------------------------------------------
##          1.f |   5.215943    2.20634     3.90   0.000      2.27654    11.95062
##          1.h |   3.513298   1.408747     3.13   0.002     1.601046    7.709499
##          cv1 |   1.197961   .0364223     5.94   0.000      1.12866    1.271518
##        _cons |   .0000347   .0000563    -6.33   0.000     1.44e-06    .0008345
## ------------------------------------------------------------------------------
## Note: _cons estimates baseline odds.
##
## . estimates store main
##
## . quietly logistic y i.f##i.h cv1
##
## . estimates store inter
##
## . lrtest main inter
##
## Likelihood-ratio test                                 LR chi2(1)  =      5.84
## (Assumption: main nested in inter)                    Prob > chi2 =    0.0157
##
## .
## . test 1.f#1.h
##
##  ( 1)  [y]1.f#1.h = 0
##
##            chi2(  1) =    5.41
##          Prob > chi2 =    0.0201
##
## .
```

Since the interaction effect is significant, we will use the `inter` model to obtain our odds ratios with confidence intervals through `lincom` (linear combination of parameters) command.

```
## . lincom 1.f, eform
##
##  ( 1)  [y]1.f = 0
##
## ------------------------------------------------------------------------------
##            y |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
## -------------+----------------------------------------------------------------
##          (1) |   20.00771   15.04885     3.98   0.000      4.58104    87.38374
## ------------------------------------------------------------------------------

## . lincom 1.f 1.f#1.h, eform
##  ( 1)  [y]1.f + [y]1.f#1.h = 0
##
## ------------------------------------------------------------------------------
##            y |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
## -------------+----------------------------------------------------------------
##          (1) |   2.581479   1.319015     1.86   0.063     .9482971    7.027367
## ------------------------------------------------------------------------------
```

Therefore,

- the OR for `f` = 1 vs. `f` = 0 when h = 0 and controlling for `cv1` is 20.1 (95% CI: 4.58, 87.4);
- the OR for `f` = 1 vs. `f` = 0 when h = 1 and controlling for `cv1` is 2.58 (95% CI: 0.95, 7.03).

```
## . lincom 1.h, eform
##
##  ( 1)  [y]1.h = 0
##
## ------------------------------------------------------------------------------
##            y |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
## -------------+----------------------------------------------------------------
##          (1) |   10.92345   7.218757     3.62   0.000     2.991185     39.8911
## ------------------------------------------------------------------------------


##  . lincom 1.h + 1.f#1.h, eform
##
##  ( 1)  [y]1.h + [y]1.f#1.h = 0
##
## ------------------------------------------------------------------------------
##            y |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
## -------------+----------------------------------------------------------------
##          (1) |   1.409389   .7762522     0.62   0.533     .4788645    4.148098
## ------------------------------------------------------------------------------
```

Therefore,

- the OR for `h` = 1 vs. `h` = 0 when f = 0 and controlling for `cv1` is 10.9 (95% CI: 2.99, 39.9);
- the OR for `h` = 1 vs. `h` = 0 when f = 1 and controlling for `cv1` is 1.41 (95% CI: 0.48, 4.15).