Interactions in Logistic Regression

Andy Grogan-Kaylor

9 Jul 2020

Background

The purpose of this tutorial is to illustrate the idea that in *logistic regression*, the β parameter for an interaction term may not accurately characterize the underlying interactive relationships.

This idea may be easier to describe if we recall the formula for a logistic regression:

$$\ln\left(\frac{P(y)}{1-P(y)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 * x_2$$

> In the above formula, the sign, and statistical significance, of β_3 may not accurately characterize the underlying relationship.

Get The Data

We start by obtaining *simulated data* from StataCorp.

```
. clear all

. graph close _all

. use http://www.stata-press.com/data/r15/margex, clear
(Artificial data for margins)
```

Describe The Data

The variables are as follows:

```
. describe
Contains data from http://www.stata-press.com/data/r15/margex.dta
  obs:         3,000                          Artificial data for margins
 vars:            11                          27 Nov 2016 14:27
 size:        78,000
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
y               float   %6.1f
outcome         byte    %2.0f
sex             byte    %6.0f      sexlbl
group           byte    %2.0f
age             float   %3.0f
distance        float   %6.2f
```

```
ycn             float   %6.1f
yc              float   %6.1f
treatment       byte    %2.0f
agegroup        byte    %8.0g       agelab
arm             byte    %8.0g
------------------------------------------------------------------------------
Sorted by: group
```

Estimate Logistic Regression

We then run a logistic regression model in which `outcome` is the dependent variable. `sex`, `age` and `group` are the independent variables. We estimate an interaction of `sex` and `age`.

We note that the regression coefficient for the interaction term is not statistically significant.

```
. logit outcome sex##c.age group
Iteration 0:   log likelihood = -1366.0718
Iteration 1:   log likelihood = -1117.9739
Iteration 2:   log likelihood = -1070.4331
Iteration 3:   log likelihood = -1068.1463
Iteration 4:   log likelihood = -1068.1394
Iteration 5:   log likelihood = -1068.1394
Logistic regression                             Number of obs     =      3,000
                                                LR chi2(4)        =     595.86
                                                Prob > chi2       =     0.0000
Log likelihood = -1068.1394                     Pseudo R2         =     0.2181

------------------------------------------------------------------------------
     outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
      female |   .6128018   .6410998     0.96   0.339    -.6437307    1.869334
         age |   .0919461    .011215     8.20   0.000     .0699652    .1139271
             |
     sex#c.age |
      female |  -.0023741   .0132292    -0.18   0.858    -.0283028    .0235547
             |
       group |  -.6267288   .1119308    -5.60   0.000    -.8461092   -.4073484
       _cons |  -5.000151   .6104382    -8.19   0.000    -6.196588   -3.803714
------------------------------------------------------------------------------
```

Margins

We use the `margins` command to estimate predicted probabilities at different values of `sex` and `age`.

```
. margins sex, at(age = (20 30 40 50 60))
Predictive margins                              Number of obs     =      3,000
Model VCE    : OIM
Expression   : Pr(outcome), predict()
1._at        : age             =          20
2._at        : age             =          30
3._at        : age             =          40
4._at        : age             =          50
5._at        : age             =          60

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _at#sex |
     1#male  |   .0147659   .0046146     3.20   0.001     .0057214    .0238104
```

1#female	.0256473	.0055867	4.59	0.000	.0146975	.0365971
2#male	.036082	.0074358	4.85	0.000	.0215081	.0506559
2#female	.0601807	.0086289	6.97	0.000	.0432683	.077093
3#male	.0850702	.009884	8.61	0.000	.0656979	.1044425
3#female	.1338511	.0108109	12.38	0.000	.1126622	.1550401
4#male	.1859699	.0163525	11.37	0.000	.1539195	.2180202
4#female	.26897	.0156965	17.14	0.000	.2382054	.2997346
5#male	.3558393	.0405971	8.77	0.000	.2762704	.4354082
5#female	.4632205	.0316107	14.65	0.000	.4012647	.5251762

Plotting Margins

`margins` provides a lot of results, which can be difficult to understand. Therefore, we use `marginsplot` to *plot* these `margins` results. The key command is `marginsplot`, which could be used on its own. I have simply added the Michigan graph scheme, as well as some options to improve the graphic design of the plot.

There certainly seems to be some kind of interaction of `sex` and `age`.

```
. marginsplot, ///
> scheme(michigan) /// michigan graph scheme
> plotopts(msize(vlarge)) /// larger plotting symbols
> plot1opts(lcolor(navy)) /// line for first group is navy
> plot2opts(lcolor(gold)) // line for second group is gold
  Variables that uniquely identify margins: age sex

. graph export mymarginsplot.png, width(500) replace
(file mymarginsplot.png written in PNG format)
```

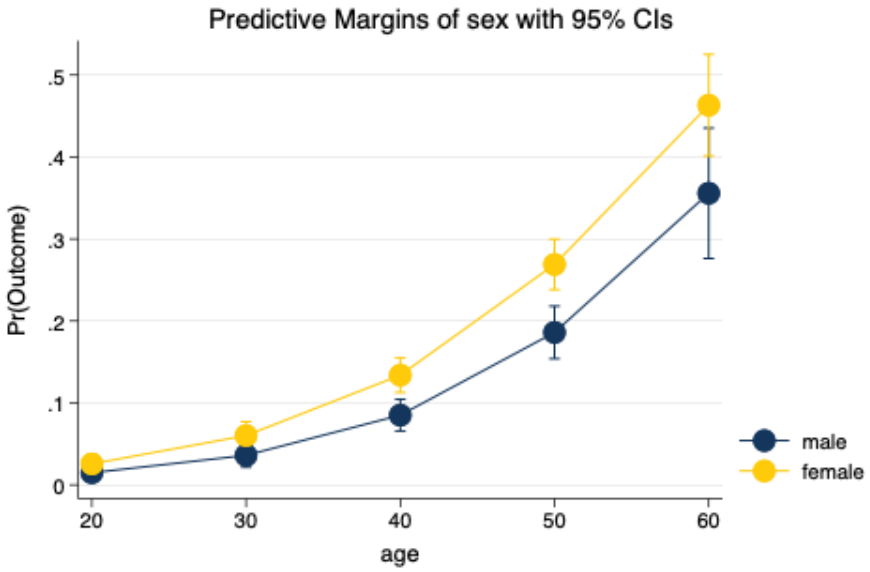

Figure 1: Margins Plot

Rerun `margins`, Posting Results

We again employ the `margins` command, this time using the `post` option so that the results of the margins command are *posted* as an estimation result. This will allow us to employ the `test` command to statistically test different margins against each other.

```
. margins sex, at(age = (20 30 40 50 60)) post
Predictive margins                              Number of obs     =      3,000
Model VCE    : OIM

Expression   : Pr(outcome), predict()
1._at        : age             =          20
2._at        : age             =          30
3._at        : age             =          40
4._at        : age             =          50
5._at        : age             =          60
```

	Delta-method Margin	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_at#sex						
1#male	.0147659	.0046146	3.20	0.001	.0057214	.0238104
1#female	.0256473	.0055867	4.59	0.000	.0146975	.0365971
2#male	.036082	.0074358	4.85	0.000	.0215081	.0506559
2#female	.0601807	.0086289	6.97	0.000	.0432683	.077093
3#male	.0850702	.009884	8.61	0.000	.0656979	.1044425
3#female	.1338511	.0108109	12.38	0.000	.1126622	.1550401
4#male	.1859699	.0163525	11.37	0.000	.1539195	.2180202
4#female	.26897	.0156965	17.14	0.000	.2382054	.2997346
5#male	.3558393	.0405971	8.77	0.000	.2762704	.4354082
5#female	.4632205	.0316107	14.65	0.000	.4012647	.5251762

`margins` with `coeflegend`

We follow up by using the `margins` command with the `coeflegend` option to see the way in which Stata has labeled the different margins.

```
. margins, coeflegend
Predictive margins                              Number of obs     =      3,000
Model VCE    : OIM

Expression   : Pr(outcome), predict()
1._at        : age             =          20
2._at        : age             =          30
3._at        : age             =          40
4._at        : age             =          50
5._at        : age             =          60
```

	Margin	Legend
_at#sex		
1#male	.0147659	_b[1bn._at#0bn.sex]
1#female	.0256473	_b[1bn._at#1.sex]
2#male	.036082	_b[2._at#0bn.sex]
2#female	.0601807	_b[2._at#1.sex]
3#male	.0850702	_b[3._at#0bn.sex]
3#female	.1338511	_b[3._at#1.sex]
4#male	.1859699	_b[4._at#0bn.sex]
4#female	.26897	_b[4._at#1.sex]
5#male	.3558393	_b[5._at#0bn.sex]
5#female	.4632205	_b[5._at#1.sex]

Testing Margins Against Each Other

Lastly, we test the margins at age 20 for men and women, and again at age 60 for men and women.

> We note that the original regression parameter for the interaction term was not statistically significant. Indeed, the margins at age 20 are not statistically significantly different by sex. However, at age 60, there is a statistically significant difference by sex.

```
. test _b[1bn._at#0bn.sex] = _b[1bn._at#1.sex] // male and female at age 20
 ( 1)  1bn._at#0bn.sex - 1bn._at#1.sex = 0
           chi2(  1) =    2.29
         Prob > chi2 =    0.1303

. test  _b[5._at#0bn.sex] =  _b[5._at#1.sex] // male and female at age 60
 ( 1)  5._at#0bn.sex - 5._at#1.sex = 0
           chi2(  1) =    4.91
         Prob > chi2 =    0.0267
```

References

Ai, C., & Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics Letters*. https://doi.org/10.1016/S0165-1765(03)00032-6

Karaca-Mandic, P., Norton, E. C., & Dowd, B. (2012). Interaction terms in nonlinear models. *Health Services Research*. https://doi.org/10.1111/j.1475-6773.2011.01314.x