Interactions in Logistic Regression

Andy Grogan-Kaylor

2024-03-09

Table of contents

1 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:

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

Warning

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

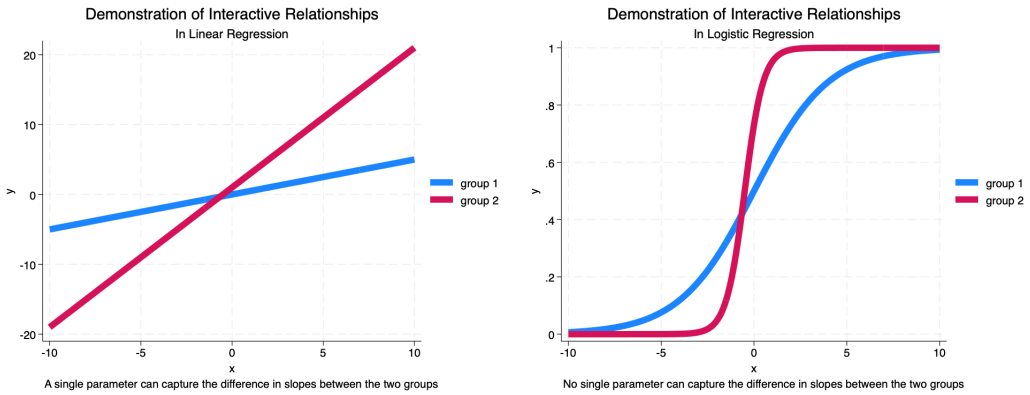

Figure 1: Demonstration of Interactive Relationships

Key Idea

In a *linear* model, a single parameter can capture the difference in slopes between the two groups. In a *non-linear* model, no single parameter can capture the difference in slopes between the two groups.

Some Calculus (Not Essential To The Discussion)

Imagine a linear model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \times x_2 + e_i$$

Here (following Ai and Norton (2003)):

$$\frac{\partial y}{\partial x_1 \partial x_2} = \beta_3$$

We use logit to describe:

$$\ln\left(\frac{P(y)}{1-P(y)}\right)$$

In the logistic model, the quantity:

$$\frac{\partial \text{logit}(y)}{\partial x_1 \partial x_2}$$

does not have such a straightforward solution, and–importantly for this discussion–is not simply equal to β_3.

2 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)
```

3 Describe The Data

The variables are as follows:

```
describe
```

```
Contains data
 Observations:             0
    Variables:             0
Sorted by:
```

4 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 i.group
```

```
no variables defined
r(111);

end of do-file
r(111);
```

5 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))
```

```
last estimates not found
r(301);

end of do-file
r(301);
```

6 Plotting Margins

`margins` provides a lot of results, which can be difficult to understand. Therefore, we use `marginsplot` to *plot* these `margins` results.

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

```
marginsplot

graph export mymarginsplot.png, width(1000) replace
```

```
previous command was not margins
r(301);

end of do-file
r(301);
```

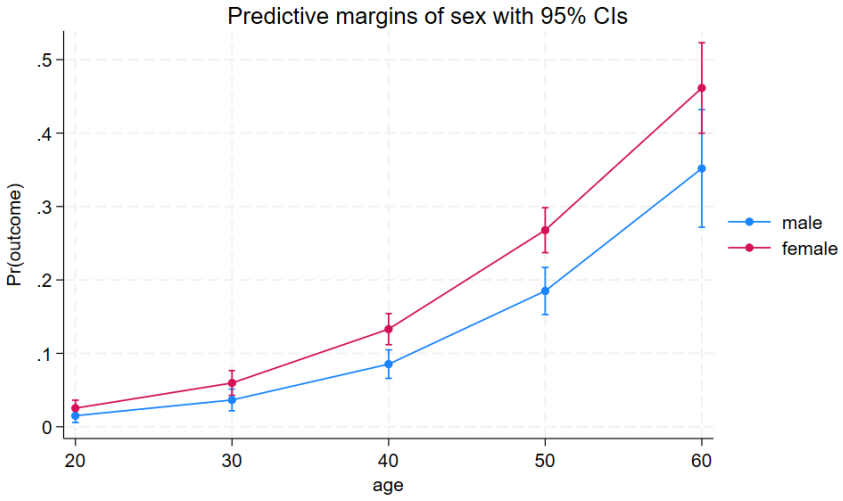

Figure 2: Margins Plot

7 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
```

```
last estimates not found
```

```
r(301);

end of do-file
r(301);
```

8 `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
```

```
last estimates not found
r(301);

end of do-file
r(301);
```

9 Testing Margins Against Each Other

Lastly, we test the margins at age 20 for men and women, and again at ages 50 and 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 ages 50 & 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

test _b[4._at#0bn.sex] = _b[4._at#1.sex] // male and female at age 50

test _b[5._at#0bn.sex] =  _b[5._at#1.sex] // male and female at age 60
```

```
last estimates not found
r(301);

end of do-file
r(301);
```

There is some suggestion that the *difference of the differences* is statistically significant. This statistical significance is only marginal [pun intended] at age 60, but truly statistically significant at age 50.

```
test _b[1bn._at#1.sex] - _b[1bn._at#0bn.sex] = _b[5._at#1.sex] - _b[5._at#0bn.sex] // test ec

test _b[1bn._at#1.sex] - _b[1bn._at#0bn.sex] = _b[4._at#1.sex] - _b[4._at#0bn.sex] // test ec
```

```
last estimates not found
r(301);

end of do-file
r(301);
```

10 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