Interactions in Logistic Regression

Andy Grogan-Kaylor

25 Mar 2020 16:25:27

The Math

$\ln(\frac{p(y)}{1-p(y)}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...$

Set $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...$ = to z.

Then

$\ln(\frac{p(y)}{1-p(y)}) = z$

$\frac{p(y)}{1-p(y)} = e^z$

$p(y) = e^z(1 - p(y))$

$p(y) = e^z - e^z(p(y))$

$e^z p(y) + p(y) = e^z$

$(1 + e^z)p(y) = e^z$

$p(y) = \frac{e^z}{1+e^z}$

$p(y) = \frac{e^{\beta_0 + \beta_1 x1 + \beta_2 x2 + ...}}{1+e^{\beta_0 + \beta_1 x1 + \beta_2 x2 + ...}}$

Simulate Some Data

```
. clear all // empty data

. set obs 10000 // set observations
number of observations (_N) was 0, now 10,000

. generate x1 = rnormal(0, 2) // normally distributed

. histogram x1, scheme(michigan)
(bin=40, start=-7.6707692, width=.36182002)

. graph export myx1graph.png, width(200) replace
(file myx1graph.png written in PNG format)
```

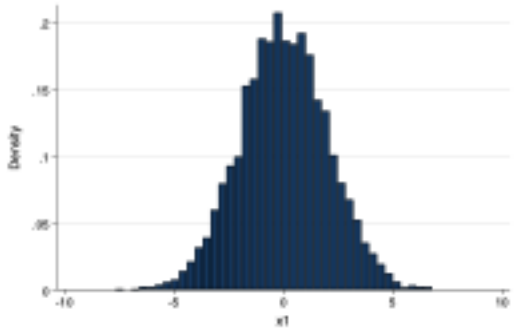

Figure 1: Histogram of x1

```
. generate x2 = rbinomial(1, .5) // categorical variable

. graph bar, over(x2) scheme(michigan)

. graph export myx2graph.png, width(200) replace
(file myx2graph.png written in PNG format)
```

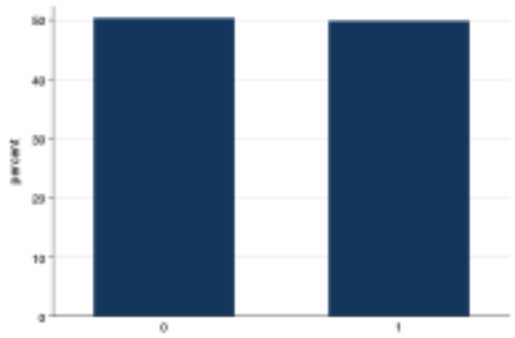

Figure 2: Bar Graph of x2

```
. summarize // descriptive statistics
```

Variable	Obs	Mean	Std. Dev.	Min	Max
x1	10,000	.0209696	1.990662	-7.670769	6.802032
x2	10,000	.4975	.5000188	0	1

Story A: Main Effects Only

Set Up The Data

```
. generate zA = x1 + x2 // first z

. generate pA = exp(zA) / (1 + exp(zA)) // probabilities

. summarize pA // descriptive statistics
```

Variable	Obs	Mean	Std. Dev.	Min	Max
pA	10,000	.5765289	.3134137	.000466	.9995913

```
. generate yA = rbinomial(1, pA) // generate y with probability p
```

i.e.

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

```
. tab yA // descriptive statistics
```

yA	Freq.	Percent	Cum.
0	4,287	42.87	42.87
1	5,713	57.13	100.00
Total	10,000	100.00	

Logistic Regression

```
. logit yA x1 x2 // does it recover the parameters?
Iteration 0:   log likelihood = -6829.4506
Iteration 1:   log likelihood = -4468.1026
Iteration 2:   log likelihood = -4417.7674
Iteration 3:   log likelihood = -4417.0613
Iteration 4:   log likelihood = -4417.0611
Logistic regression                             Number of obs     =     10,000
                                                LR chi2(2)        =    4824.78
                                                Prob > chi2       =     0.0000
Log likelihood = -4417.0611                     Pseudo R2         =     0.3532
```

yA	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
x1	1.024776	.0210616	48.66	0.000	.983496	1.066056
x2	1.072611	.0543806	19.72	0.000	.9660266	1.179194
_cons	-.0642869	.0366253	-1.76	0.079	-.1360711	.0074974

```
. predict yhatA // predicted probabilities
(option pr assumed; Pr(yA))
```

Story B: Main Effects + Interactions

Set Up The Data

```
. generate zB = x1 + x2 + (.75 * x1 * x2) // second z

. generate pB = exp(zB) / (1 + exp(zB)) // probabilities

. summarize pB // descriptive statistics
```

Variable	Obs	Mean	Std. Dev.	Min	Max
pB	10,000	.5532364	.3490571	.0000253	.9999975

```
. generate yB = rbinomial(1, pB) // generate y with probability p
```

i.e.

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

```
. tab yB // descriptive statistics
```

yB	Freq.	Percent	Cum.
0	4,421	44.21	44.21
1	5,579	55.79	100.00
Total	10,000	100.00	

Logistic Regression

```
. logit yB c.x1##i.x2 // does it recover the parameters?
Iteration 0:   log likelihood = -6864.2729
Iteration 1:   log likelihood =  -4010.245
Iteration 2:   log likelihood = -3918.6259
Iteration 3:   log likelihood = -3913.3196
Iteration 4:   log likelihood =  -3913.308
Iteration 5:   log likelihood =  -3913.308
Logistic regression                             Number of obs     =     10,000
                                                LR chi2(3)        =    5901.93
                                                Prob > chi2       =     0.0000
Log likelihood =  -3913.308                     Pseudo R2         =     0.4299
```

yB	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
x1	.9739211	.0279227	34.88	0.000	.9191937	1.028649
1.x2	1.0046	.0626434	16.04	0.000	.8818216	1.127379
x2#c.x1						
1	.7762292	.0584557	13.28	0.000	.6616582	.8908003
_cons	.0350799	.0360535	0.97	0.331	-.0355837	.1057434

```
. predict yhatB // predicted probabilities
(option pr assumed; Pr(yB))
```

Inspect The Situation With A Graph

Think for a moment about what an interaction term $\beta x_1 * x_2$ is *supposed* to capture: the *difference* between slopes.

Notice that at different values of x the *difference* between the two slopes is *different*

At $x = 0$, the **orange** line is steeper than the **green** line.

At $x = 2$, the **green** line is steeper than the **orange** line.

No single static parameter can capture this changing difference between two slopes.

```
. twoway ///
> (scatter yB x1 if x2 == 0, msize(tiny)) /// points
> (scatter yB x1 if x2 == 1, msize(tiny)) /// points
> (scatter yhatB x1 if x2 == 0, msize(tiny)) ///
> (scatter yhatB x1 if x2 == 1, msize(tiny)), ///
> xline(0 2) ///
> title("Logit Curves for 2 Groups") ///
> sub("Model With Interaction") ///
> scheme(michigan)

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

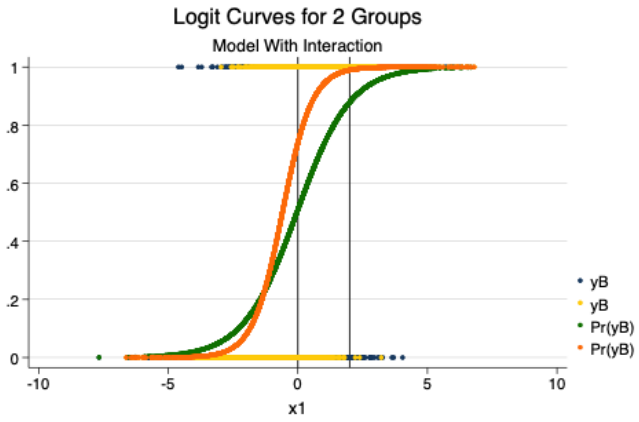

Figure 3: Logistic Regression With Interactions