Logistic Regression With Covariates

Andy Grogan-Kaylor

9 Sep 2020 09:58:09

Background

In linear regression, interpretation of coefficients is *somewhat* straightforward. We might first estimate:

$y = \beta_0 + \beta_1 x_1 + e_i$

and then:

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

and would say–in the second equation–that β_1 is an estimate that accounts for the association of x_2 and y.

However, in logistic regression, the situation is somewhat different.

As Allison (1999) notes:

> Unfortunately, there is a potential pitfall in cross-group comparisons of logit or probit coefficients that has largely gone unnoticed. Unlike linear regression coefficients, coefficients in these binary regression models are confounded with residual variation (unobserved heterogeneity). Differences in the degree of residual variation across groups can produce apparent differences in coefficients that are not indicative of true differences in causal effects.

While the mathematics of this relationship are somewhat difficult–though clearly presented in Allison's (1999) article–the finding can be easily seen in simulated data.

Simulate Data

```
. clear all

. cd "/Users/agrogan/Desktop/newstuff/categorical/logistic-and-covariates"
/Users/agrogan/Desktop/newstuff/categorical/logistic-and-covariates

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

. set seed 3846 // random seed

. generate x1 = rnormal() // normally distributed x

. histogram x1, scheme(michigan)
(bin=40, start=-3.7857256, width=.19587822)

. graph export histogram1.png, width(500) replace
(file histogram1.png written in PNG format)
```

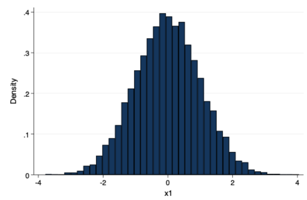

Figure 1: Histogram of x1

```
. generate x2 = rnormal() // normally distributed z

. histogram x2, scheme(michigan)
(bin=40, start=-3.9428685, width=.19152238)

. graph export histogram2.png, width(500) replace
(file histogram2.png written in PNG format)
```

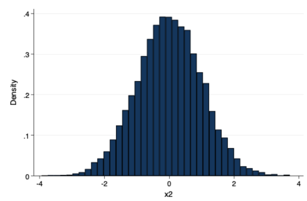

Figure 2: Histogram of x2

```
. generate e = rnormal(0, .5) // normally distributed error
```

Since they were generated independently, x_1 and x_2 are relatively uncorrelated.

```
. corr x1 x2 // x1 and x2 are uncorrelated
(obs=10,000)

             |       x1       x2
-------------+------------------
          x1 |   1.0000
          x2 |   0.0150   1.0000

. generate y1 = x1 + x2 + e // dependent variable
```

Linear Regression

```
. regress y1 x1
      Source |       SS           df       MS      Number of obs   =    10,000
-------------+----------------------------------   F(1, 9998)      =   8571.07
       Model |   10888.525         1   10888.525   Prob > F        =    0.0000
    Residual |  12701.2625     9,998  1.27038033   R-squared       =    0.4616
-------------+----------------------------------   Adj R-squared   =    0.4615
       Total |  23589.7876     9,999  2.35921468   Root MSE        =    1.1271

------------------------------------------------------------------------------
          y1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   1.024698   .0110682    92.58   0.000     1.003002    1.046394
       _cons |   .0013059   .0112712     0.12   0.908     -.020788    .0233997
------------------------------------------------------------------------------

. est store OLS1 // store estimates
```

```
. regress y1 x1 x2
      Source |       SS           df       MS      Number of obs   =    10,000
-------------+----------------------------------   F(2, 9997)      =  41868.07
       Model |  21073.8459         2  10536.9229   Prob > F        =    0.0000
    Residual |  2515.94171     9,997  .251669672   R-squared       =    0.8933
-------------+----------------------------------   Adj R-squared   =    0.8933
       Total |  23589.7876     9,999  2.35921468   Root MSE        =    .50167

------------------------------------------------------------------------------
          y1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   1.009826   .0049269   204.96   0.000     1.000169    1.019484
          x2 |   1.006154   .0050014   201.17   0.000     .9963505    1.015958
       _cons |   .0015213   .0050167     0.30   0.762    -.0083125     .011355
------------------------------------------------------------------------------

. est store OLS2 // store estimates
```

Note that the coefficients for x_1 in the two models are relatively close.

```
. estimates table OLS1 OLS2, b(%7.4f) star // table comparing estimates

----------------------------------------
    Variable |    OLS1          OLS2
-------------+--------------------------
          x1 |  1.0247***     1.0098***
          x2 |                1.0062***
       _cons |  0.0013        0.0015
----------------------------------------
legend: * p<0.05; ** p<0.01; *** p<0.001
```

Logistic Regression

```
. generate prob_y2 = exp(x1 + x2 + e) / (1 + exp(x1 + x2 + e)) // dependent variable

. recode prob_y2 (0/.5 =0)(.5/1 = 1), generate(y2) // recode probabilites as observed val
> ues
(10000 differences between prob_y2 and y2)

. logit y2 x1
Iteration 0:   log likelihood = -6931.3566
Iteration 1:   log likelihood = -5193.5531
Iteration 2:   log likelihood = -5191.3673
Iteration 3:   log likelihood = -5191.3654
Iteration 4:   log likelihood = -5191.3654
Logistic regression                             Number of obs     =     10,000
                                                LR chi2(1)        =    3479.98
                                                Prob > chi2       =     0.0000
Log likelihood = -5191.3654                     Pseudo R2         =     0.2510

------------------------------------------------------------------------------
          y2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   1.529607   .0329772    46.38   0.000     1.464973    1.594241
       _cons |   .0205374   .0240145     0.86   0.392    -.0265302     .067605
------------------------------------------------------------------------------

. est store logit1

. logit y2 x1 x2
Iteration 0:   log likelihood = -6931.3566
Iteration 1:   log likelihood = -2326.0511
Iteration 2:   log likelihood = -2285.4234
Iteration 3:   log likelihood = -2285.2877
Iteration 4:   log likelihood = -2285.2877
Logistic regression                             Number of obs     =     10,000
```

```
                                                Prob > chi2       =     0.0000
                                                LR chi2(2)        =    9292.14
                                                Prob > chi2       =     0.0000
Log likelihood = -2285.2877                     Pseudo R2         =     0.6703
```

y2	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
x1	3.694725	.0867616	42.58	0.000	3.524675	3.864774
x2	3.716715	.0876762	42.39	0.000	3.544873	3.888557
_cons	.0369852	.0375883	0.98	0.325	-.0366864	.1106569

```
Note: 6 failures and 4 successes completely determined.

. est store logit2
```

Note that the coefficients for x_1 in the two models are rather different, even though x_1 and x_2 are, by definition, uncorrelated.

```
. estimates table logit1 logit2, b(%7.4f) star // table comparing estimates
```

Variable	logit1	logit2
x1	1.5296***	3.6947***
x2		3.7167***
_cons	0.0205	0.0370

```
legend: * p<0.05; ** p<0.01; *** p<0.001
```

References

Allison, P. D. (1999). Comparing logit and probit coefficients across groups. *Sociological Methods and Research*. https://doi.org/10.1177/0049124199028002003