Logistic Regression With Covariates

Andy Grogan-Kaylor

8 Sep 2020 16:08:06

Background

Simulate Data

```
. clear all

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

. set obs 1000
number of observations (_N) was 0, now 1,000

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

. histogram x1, scheme(michigan)
(bin=29, start=-3.1592772, width=.22074086)

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

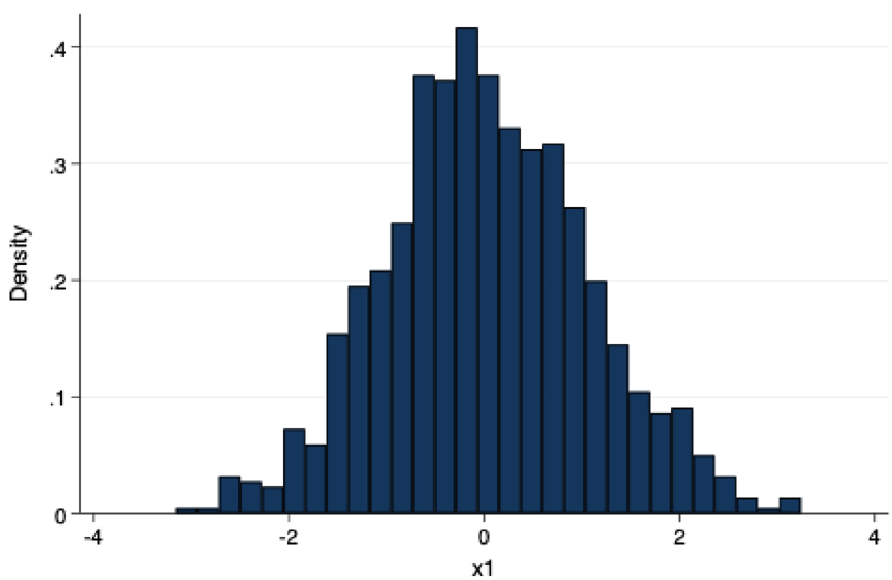

Figure 1: Histogram of x1

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

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

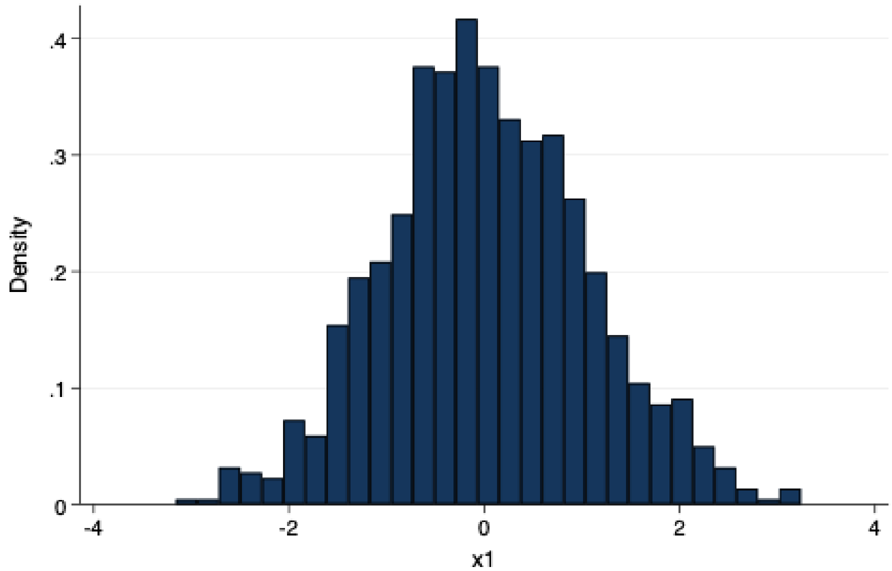

Figure 2: Histogram of x2

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

. corr x1 x2 // x1 and x2 are uncorrelated
(obs=1,000)

             |       x1       x2
-------------+------------------
          x1 |   1.0000
          x2 |   0.0596   1.0000

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

Linear Regression

```
. regress y1 x1

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(1, 998)       =    635.15
       Model |  1289.80971         1  1289.80971   Prob > F        =    0.0000
    Residual |  2026.66544       998  2.03072689   R-squared       =    0.3889
-------------+----------------------------------   Adj R-squared   =    0.3883
       Total |  3316.47515       999  3.31979494   Root MSE        =     1.425

------------------------------------------------------------------------------
          y1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   1.088659   .0431971    25.20   0.000     1.003892    1.173427
       _cons |   .0278531   .0450746     0.62   0.537    -.0605987    .1163049
------------------------------------------------------------------------------

. est store OLS1 // store estimates

. regress y1 x1 x2
```

```
      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(2, 997)       =   1153.65
       Model |   2315.8002         2   1157.9001   Prob > F        =    0.0000
    Residual |  1000.67495       997    1.003686   R-squared       =    0.6983
-------------+----------------------------------   Adj R-squared   =    0.6977
       Total |  3316.47515       999  3.31979494   Root MSE        =    1.0018

------------------------------------------------------------------------------
          y1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   1.030682   .0304229    33.88   0.000     .9709817    1.090382
          x2 |   .9874388   .0308843    31.97   0.000     .9268331    1.048044
       _cons |   .0081816   .0316947     0.26   0.796    -.0540144    .0703775
------------------------------------------------------------------------------

. est store OLS2 // store estimates

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

----------------------------------------
    Variable |    OLS1          OLS2
-------------+--------------------------
          x1 |  1.0887***     1.0307***
          x2 |                0.9874***
       _cons |  0.0279        0.0082
----------------------------------------
legend: * p<0.05; ** p<0.01; *** p<0.001
```

Logistic Regression

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

. recode prob_y2 (0/.5 =0)(.5/1 = 1), generate(y2) // recode probabilites as observed val
> ues
(1000 differences between prob_y2 and y2)

. logit y2 x1
Iteration 0:   log likelihood = -693.11518
Iteration 1:   log likelihood = -550.43417
Iteration 2:   log likelihood = -550.34901
Iteration 3:   log likelihood = -550.34899
Logistic regression                             Number of obs     =      1,000
                                                LR chi2(1)        =     285.53
                                                Prob > chi2       =     0.0000
Log likelihood = -550.34899                     Pseudo R2         =     0.2060

------------------------------------------------------------------------------
          y2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   1.282626   .0926296    13.85   0.000     1.101075    1.464177
       _cons |   .0044323   .0733194     0.06   0.952     -.139271    .1481356
------------------------------------------------------------------------------

. est store logit1

. logit y2 x1 x2
Iteration 0:   log likelihood = -693.11518
Iteration 1:   log likelihood = -399.88043
Iteration 2:   log likelihood = -399.52919
Iteration 3:   log likelihood = -399.52837
Iteration 4:   log likelihood = -399.52837
Logistic regression                             Number of obs     =      1,000
                                                LR chi2(2)        =     587.17
                                                Prob > chi2       =     0.0000
Log likelihood = -399.52837                     Pseudo R2         =     0.4236

------------------------------------------------------------------------------
          y2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
             |
```

```
          x1 |    1.80266   .1291406    13.96   0.000     1.549549    2.055771
          x2 |   1.644651   .1215883    13.53   0.000     1.406342     1.88296
       _cons |   -.060496   .0882002    -0.69   0.493    -.2333652    .1123732
------------------------------------------------------------------------------

. est store logit2

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

----------------------------------------
    Variable |   logit1       logit2    
-------------+--------------------------
          x1 |  1.2826***    1.8027***  
          x2 |               1.6447***  
       _cons |  0.0044      -0.0605     
----------------------------------------
legend: * p<0.05; ** p<0.01; *** p<0.001
```

References

I've been inspired in this disussion by Jonathan Bartlett's discussion of these issues: https://thestatsgeek.com/2017/05/11/odds-ratios-collapsibility-marginal-vs-conditional-gee-vs-glmms/