Logistic Regression The Basics

Andy Grogan-Kaylor

3 May 2021 09:46:51

Logistic Regression

Basic handout on logistic regression for a binary dependent variable.

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
 Observations:         3,000                  Artificial data for margins
    Variables:            11                  27 Nov 2016 14:27
-------------------------------------------------------------------------------
Variable      Storage   Display    Value
    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
```

The Equation

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

Here $p(outcome)$ is the probability of the outcome.

$\frac{p(outcome)}{1-p(outcome)}$ is the *odds* of the outcome.

Hence, $\ln\left(\frac{p(outcome)}{1-p(outcome)}\right)$ is the *log odds*.

Logistic regression returns a β coefficient for each independent variable x.

These β coefficients can then be *exponentiated* to obtain *odds ratios*: $OR = e^{\beta}$

Estimate Logistic Regression (`logit y x`)

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

```
. logit outcome i.sex c.age i.group
Iteration 0:   log likelihood = -1366.0718
Iteration 1:   log likelihood = -1111.4595
Iteration 2:   log likelihood =  -1069.588
Iteration 3:   log likelihood =      -1068
Iteration 4:   log likelihood = -1067.9941
Iteration 5:   log likelihood = -1067.9941
Logistic regression                                     Number of obs =  3,000
                                                        LR chi2(4)    = 596.16
                                                        Prob > chi2   = 0.0000
Log likelihood = -1067.9941                             Pseudo R2     = 0.2182
------------------------------------------------------------------------------
     outcome | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         sex |
     female  |   .4991622   .1347463     3.70   0.000     .2350643      .76326
         age |   .0902429   .0064801    13.93   0.000     .0775421    .1029437
             |
       group |
          2  |  -.5855242   .1350192    -4.34   0.000     -.850157   -.3208915
          3  |  -1.360208   .2914263    -4.67   0.000    -1.931393   -.7890228
             |
       _cons |  -5.553038   .3498204   -15.87   0.000    -6.238674   -4.867403
------------------------------------------------------------------------------
```

Odds Ratios (`logit y x, or`)

We re-run the model with exponentiated coefficients (e^{β} to obtain odds ratios.

```
. logit outcome i.sex c.age i.group, or
Iteration 0:   log likelihood = -1366.0718
Iteration 1:   log likelihood = -1111.4595
Iteration 2:   log likelihood =  -1069.588
Iteration 3:   log likelihood =      -1068
Iteration 4:   log likelihood = -1067.9941
Iteration 5:   log likelihood = -1067.9941
Logistic regression                                     Number of obs =  3,000
                                                        LR chi2(4)    = 596.16
                                                        Prob > chi2   = 0.0000
```

```
Log likelihood = -1067.9941                             Pseudo R2     = 0.2182

------------------------------------------------------------------------------
     outcome | Odds ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         sex |
     female  |    1.64734    .221973     3.70   0.000      1.26499    2.145258
         age |    1.09444   .0070921    13.93   0.000     1.080628    1.108429
             |
       group |
          2  |   .5568139   .0751806    -4.34   0.000     .4273478     .725502
          3  |   .2566074   .0747822    -4.67   0.000     .1449462    .4542885
             |
       _cons |   .0038757   .0013558   -15.87   0.000     .0019524    .0076933
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

β Coefficients and Odds Ratios

Substantively	β	OR
x is associated with an increase in y	> 0.0	> 1.0
no association	0.0	1.0
x is associated with a descrease in y	< 0.0	< 1.0

Coefficients, Standard Errors, p values, and Confidence Intervals

- z statistic: $z = \frac{\beta}{se}$.
- p value if $z_{\text{observed}} > 1.96$ then $p < .05$.
- CI $= \beta \pm 1.96 * se$

Hence for the coefficient for `sex`, the confidence interval is:

$$.4991622 \pm (1.959964 * .1347463) = (.2350643, .7632601)$$

Confidence intervals for *odds ratios* (e^{β}) are obtained by exponentiating the confidence interval for the β coefficients. As a result of this non-linear transformation, confidence intervals for odds ratios are not symmetric.