Week 7. Logistic Regression 1

When to use

1. Use Logistic Regression when your outcome variable (= dependent variable) is a dummy variable.
   E.g. employment = 0, unemployment = 1.

2. Independent variables can be any variables.

In a Nutshell

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \sum_{k=0}^{K} \beta_k X_{ik} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_K X_{iK}$$

1. $\left(\frac{p_i}{1-p_i}\right)$ is called odds. For example, if the probability of employment is 90%, $\left(\frac{.90}{1-.90}\right) = 9$. The odds of employment is 9.

2. The odds is log transformed.

3. Thus,

$$\exp\left(\ln\left(\frac{p_i}{1-p_i}\right)\right) = \frac{p_i}{1-p_i} = \mathrm{e}^{\beta_0+\beta_1 X_{i1}+\cdots+\beta_K X_{iK}} = \mathrm{e}^{\beta_0}\mathrm{e}^{\beta_1 X_{i1}} \cdots \mathrm{e}^{\beta_K X_{iK}}$$

4. The interpretation is exactly the same as OLS with a log-transformed dependent variable. In OLS, y is log transformed. In Logistic Regression, Logit = $\left(\frac{p_i}{1-p_i}\right)$ is log transformed.

5. "exp(β_k)" is odds ratio, which quantifies how much "times" of odds increases when X increases by 1 unit compared to the reference group.

6. Interpretation 1: As X increases by 1 unit, the odds increase by exp(β_k) times compared to the reference point. Odds ratio is a ratio of two odds.

7. Interpretation 2: As X increases by 1 unit, the log odds (= Logit) increases by β_k.

8. For example, let's say that the probabilities of employment are 90% for BA+ and 80% for HSG. Odds of employment for BA+ is 9 (= .9/(1-.9)), and odds of employment for HSG is 4 (= .8/(1-.8). Thus odds ratio is 9/4 = 2.25. Compared to HSG, the odds of employment is 2.25 times higher for BA+.

$$\ln\left(\frac{p_i}{1-p_i}\right) = 1.39 + .81BA$$

   Probability of employment for HSG = .8
   Odds for HSG = .8/(1-.8) = 4
   Log of odds = Logit for HSG = ln(4) = 1.39, which is the constant in the the above equation.

Probability of employment for BA = .9
Odds for BA = .9/(1-.9) = 9
Log of odds = Logit for BA = ln(9) = 2.2, which is 1.39+.81 in the the above equation.
Odds ratio = $\left(\frac{.9}{1-.9}\right) \div \left(\frac{.8}{1-.8}\right) = \frac{.9(1-.8)}{.8(1-.9)} = 2.25$

If the odds of employment are the same between BA and HSG, the odds ratio will be 1. In this case, the coefficient of logistic regression is 0 ($\ln(1) = 0, exp(0) = 1$).

9. Because $\left(\frac{p_i}{1-p_i}\right) = \exp(\sum_{k=0}^{K} \beta_k X_{ik})$,
$p_i = \frac{\exp(\sum_{k=0}^{K} \beta_k X_{ik})}{1 + \exp(\sum_{k=0}^{K} \beta_k X_{ik})}$

For HSG, $\exp(1.39)/(1 + \exp(1.39)) = .80$
For BA, $\exp(2.2)/(1 + \exp(2.2)) = .90$

10. Statistical significance. Interpret the same as OLS.
11. Model fitness. Report -2LL (log likelihood times -2) or LL. Stata will also provide pseudo r-squared.

Logit

1. Why Logit? Why not OLS? A binary outcome is either 1 or 0. The probability of an event cannot go higher than 1 and lower than 0. The probability distribution should look like:

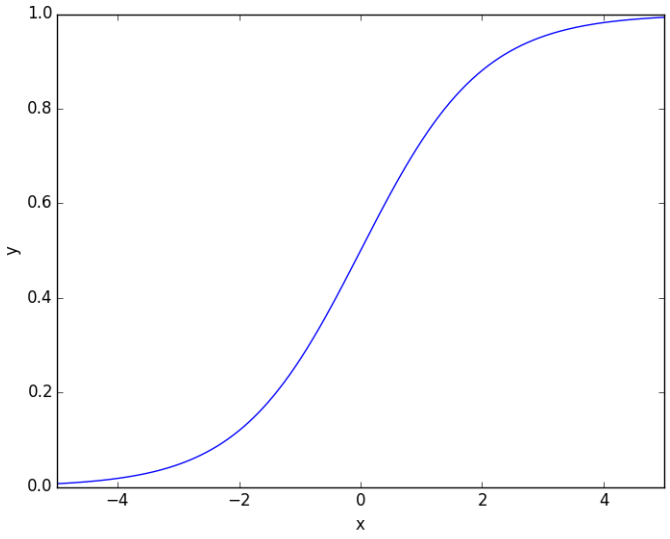

Figure 1: Probability Distribution

The expected value of OLS can go outside the 0 to 1 range. Unlike OLS, the expected probability of Logit is ranged always between 0 and 1.

2. Logit = log odds, $\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right)$
3. As the probability goes down to zero the odds approach zero and the logit approaches $-\infty$.
4. At the other extreme, as the probability approaches one the odds approach $+\infty$ and so does the logit.
5. Thus, logits map probabilities from the (0,1) to the entire real line $(-\infty, +\infty)$.
6. Note that if p = .50, the odds are even and the logit is zero. Negative logits represent probabilities below one half and positive logits correspond to probabilities above one half.
7. Because logits transfer the range from (0,1) to $(-\infty, +\infty)$, regression analyses become available in logistic regression. That is, we assume the logit of probability p_i (not probability itself) follows a linear model.
8. That is,

LPM: Linear Probability Model

1. Estimate OLS with a dummy dependent variable.
2. From the previous example, you will get the following result:

$$y = .80 + .10BA + e$$

   (a) Note that unlike logistic regression, LPM assumes that probability (p_i) itself follows a linear model.

   (b) Apparently, however, probability (p_i) is not linear. There must be ceiling and floor effects.

3. Problems of LPM

   (a) The predicted outcome can be outside the range of 0 to 1.

   (b) The dependent variable, y_i, for group i is distributed as binomial, with variance $n_i p_i(1-p_i)$, whereas the sample proportion p_i has variance of $p_i(1-p_i)/n_i$. Thus, the variance in the dependent variable depends on the size of the group and the probability of success (=1). Since these quantities are not constant across groups, the errors are heteroscedastic. As a result, the standard errors are not correct.

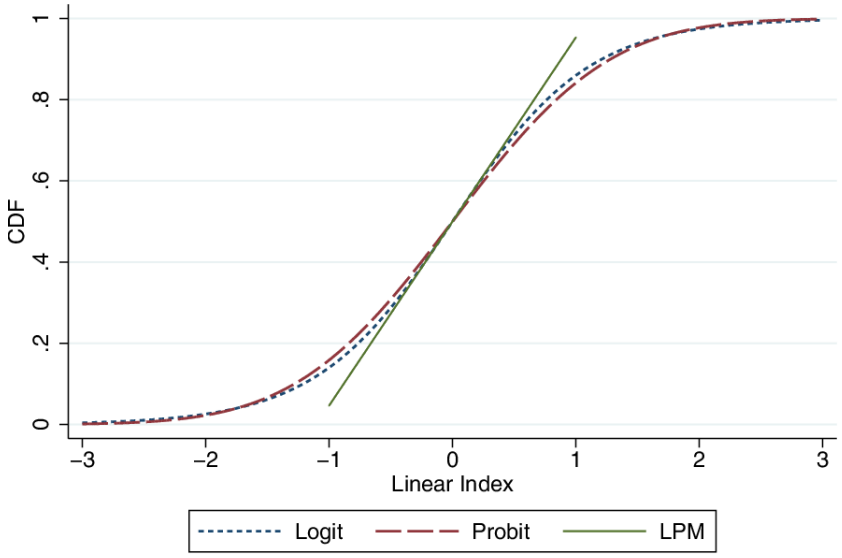

Figure 2: Logit vs. LPM

4. Solutions

   (a) For the 1st problem, unless you tries to compute the extreme cases, LPM does not have a problem of (a).

   (b) For the 2nd problem, apply the weight, $w_i = n_i/(p_i(1-p_i)$. This is a weighted least square model. Thus, OLS will minimize the weighted sum of squares (or weighted error squares) instead of minimizing the error squared. Simply put, use "robust" option in Stata. Recall that as long as you apply weight (=pw), Stata automatically applies the robust option.

Logit: Stata Results

$$\text{Logit}(p^{emp}) = \alpha + \sum \beta Edu_j + \gamma Female + \delta age + \pi age^2$$

where p^{emp} is probability of employment. In the following logistic regression, the dependent variable is emp (1 = employed, 0 = unemployed).

```
. tab emp [aw=perwt]

        emp |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 | 1,115.0065        5.81        5.81
          1 | 18,059.993       94.19      100.00
------------+-----------------------------------
      Total |     19,175      100.00


. logit emp i.edu female age age2 [pw=perwt]

Iteration 0:   log pseudolikelihood = -145438.94
Iteration 1:   log pseudolikelihood = -140763.89
Iteration 2:   log pseudolikelihood = -140156.23
Iteration 3:   log pseudolikelihood = -140155.26
Iteration 4:   log pseudolikelihood = -140155.26

Logistic regression                               Number of obs   =      19175
                                                  Wald chi2(7)    =     200.72
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -140155.26                 Pseudo R2       =     0.0363

------------------------------------------------------------------------------
             |               Robust
         emp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         edu |
          2  |   .4429065   .1249446     3.54   0.000     .1980195    .6877934
          3  |   .6875269   .1237244     5.56   0.000     .4450316    .9300223
          4  |   1.431148   .1363582    10.50   0.000     1.163891    1.698405
          5  |   1.887059   .1724174    10.94   0.000     1.549127    2.224991
             |
      female |  -.2252276   .0792752    -2.84   0.004     -.380604   -.0698511
         age |   .6173922   .4198824     1.47   0.141    -.2055622    1.440347
        age2 |  -.0077081   .0053011    -1.45   0.146     -.018098    .0026818
       _cons |  -10.21442   8.282906    -1.23   0.218    -26.44862    6.019779
------------------------------------------------------------------------------
```

To get the odds ratio, do the following:

```
. logit, or

Logistic regression                                     Number of obs   =      19175
                                                        Wald chi2(7)    =     200.72
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -140155.26                       Pseudo R2       =     0.0363

------------------------------------------------------------------------------
             |               Robust
         emp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         edu |
          2  |   1.557227   .1945671     3.54   0.000     1.218986    1.989321
          3  |   1.988791    .246062     5.56   0.000     1.560539    2.534566
          4  |   4.183499   .5704543    10.50   0.000     3.202369    5.465224
          5  |   6.599929   1.137943    10.94   0.000     4.707358    9.253398
             |
      female |   .7983345   .0632881    -2.84   0.004     .6834485    .9325327
         age |   1.854087   .7784984     1.47   0.141     .8141895    4.222159
        age2 |   .9923215   .0052604    -1.45   0.146     .9820648    1.002685
       _cons |   .0000366   .0003035    -1.23   0.218     3.26e-12    411.4875
------------------------------------------------------------------------------
```

Note that `logistic emp i.edu female age age2 [pw=perwt]` will report the identical results with odds ratio.

"margins" command

```
. margins

Predictive margins                                      Number of obs   =      19175
Model VCE    : Robust

Expression   : Linear prediction, predict()

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |    .941851   .0021221   443.84   0.000     .9376916    .9460105
------------------------------------------------------------------------------


. margins i.edu

Predictive margins                                      Number of obs   =      19175
Model VCE    : Robust
```

```
Expression   : Linear prediction, predict()

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         edu |
          1  |   .8796635   .0105679    83.24   0.000     .8589494    .9003775
          2  |   .9184215   .0054778   167.66   0.000     .9076845    .9291585
          3  |   .9347527   .0043027   217.25   0.000     .9263191    .9431862
          4  |   .9679397   .0028386   340.99   0.000     .9623758    .9735036
          5  |   .9796876   .0028365   345.39   0.000     .9741279    .9852473
------------------------------------------------------------------------------



. margins, dydx(i.edu)

Average marginal effects                          Number of obs   =      19175
Model VCE    : Robust

Expression   : Linear prediction, predict()
dy/dx w.r.t. : 2.edu 3.edu 4.edu 5.edu

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         edu |
          2  |    .038758   .0118878     3.26   0.001      .015457    .0620591
          3  |   .0550892   .0114168     4.83   0.000     .0327112    .0774672
          4  |   .0882762    .010942     8.07   0.000     .0668289    .1097236
          5  |   .1000242   .0109478     9.14   0.000     .0785655    .1214828
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.



. margins, at(age=40)

Predictive margins                                Number of obs   =      19175
Model VCE    : Robust

Expression   : Linear prediction, predict()
at           : age             =          40

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .9570455   .0106589    89.79   0.000     .9361531    .9779379
------------------------------------------------------------------------------
```

```
. margins, dydx(age)

Average marginal effects                          Number of obs   =      19175
Model VCE    : Robust

Expression   : Linear prediction, predict()
dy/dx w.r.t. : age

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0333868   .0232608     1.44   0.151    -.0122065    .0789801
------------------------------------------------------------------------------
```

LPM: Stata Results

Compare the results below with the results with "margins" command.

```
. reg emp i.edu female age age2 [pw=perwt]
(sum of wgt is   6.5559e+05)

Linear regression                                      Number of obs =   19175
                                                       F(  7, 19167) =   29.93
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0158
                                                       Root MSE      =  .23222

------------------------------------------------------------------------------
             |               Robust
         emp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         edu |
          2  |    .038758   .0118878     3.26   0.001      .015457    .0620591
          3  |   .0550892   .0114168     4.83   0.000     .0327112    .0774672
          4  |   .0882762    .010942     8.07   0.000     .0668289    .1097236
          5  |   .1000242   .0109478     9.14   0.000     .0785655    .1214828
             |
      female |  -.0120728   .0042987    -2.81   0.005    -.0204986   -.0036471
         age |   .0333868   .0232608     1.44   0.151    -.0122065    .0789801
        age2 |  -.0004171   .0002935    -1.42   0.155    -.0009923    .0001582
       _cons |   .2203937   .4597176     0.48   0.632    -.6806931    1.121481
------------------------------------------------------------------------------
```