An Introduction to Logistic and Probit Regression Models

Chelsea Moore

Goals

- Brief overview of logistic and probit models
- Example in Stata
- Interpretation within & between models

Binary Outcome

- Examples:
  - Yes/No
  - Success/Failure
  - Heart Attack/No Heart Attack
  - In/Out of the Labor Force

Modeling a Binary Outcome

- Latent Variable Approach
  - We can think of y* as the underlying latent propensity that y=1
    - Example 1: For the binary variable, heart attack/no heart attack, y* is the propensity for a heart attack.
    - Example 2: For the binary variable, in/out of the labor force, y* is the propensity to be in the labor force.

$$y^* = \alpha + \beta x + \varepsilon$$

$$y_i = \begin{cases} 1 & if\ y_i^* > \tau \\ 0 & if\ y_i^* \leq \tau \end{cases}$$

Where τ is the threshold

Logit versus Probit

- Since y* is unobserved, we use do not know the distribution of the errors, ε

- In order to use maximum likelihood estimation (ML), we need to make some assumption about the distribution of the errors.

Logit versus Probit

- The difference between Logistic and Probit models lies in this assumption about the distribution of the errors

- Logit

$$\ln\left(\frac{p_i}{(1-p_i)}\right) = \sum_{k=0}^{k=n} \beta_k x_{ik}$$

  - *Standard logistic* distribution of errors

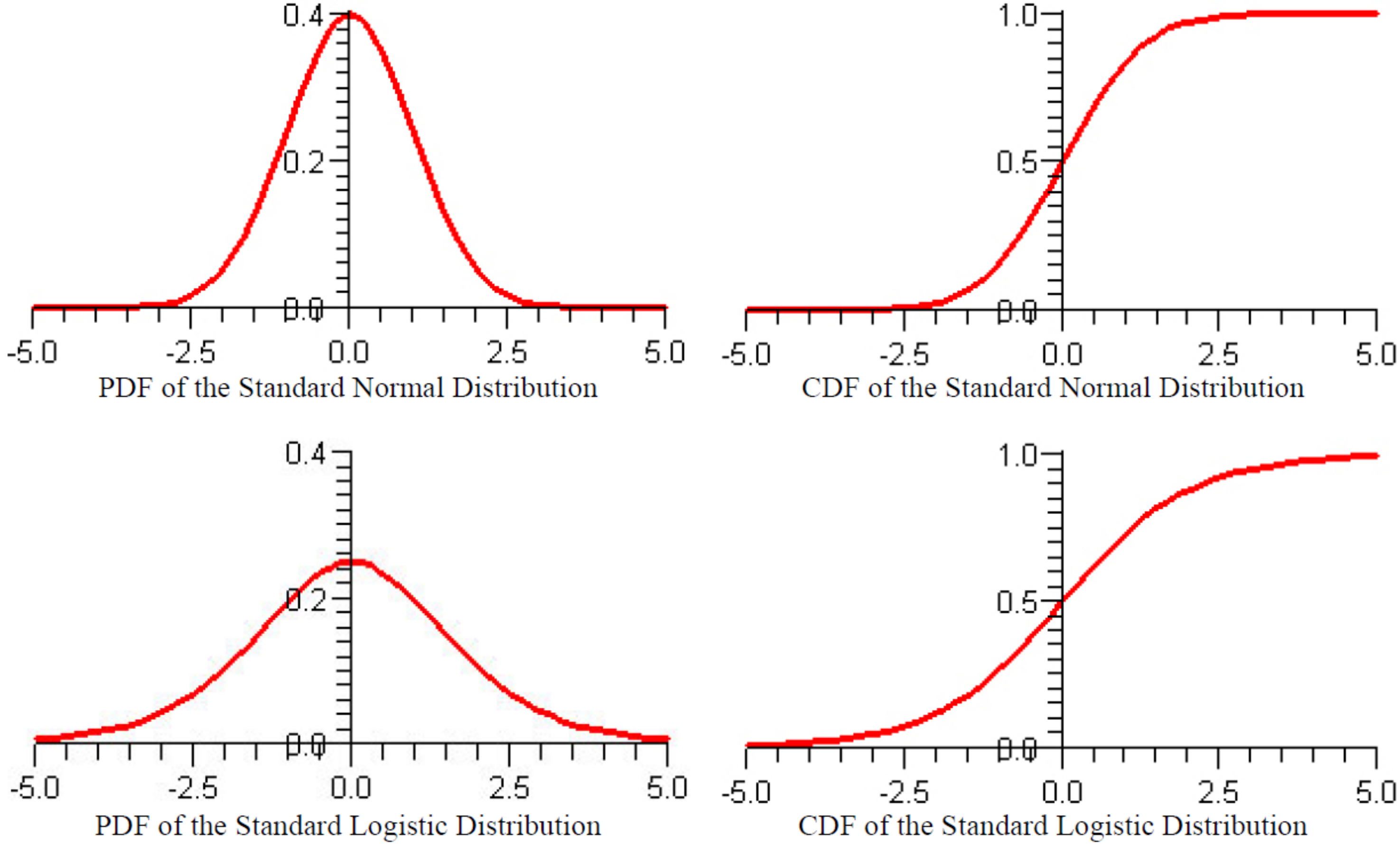
- Probit

$$\Phi^{-1}(p_i) = \sum_{k=0}^{k=n} \beta_k x_{ik}$$

  - *Normal* distribution of errors

Probability Density Function (PDF)and Cumulative Distribution Function (CDF)

Figure 1.1 The Standard Normal and Standard Logistic Probability Distributions



Source: Park (2010)

Which to choose?

- Results tend to be very similar
- Preference for one over the other tends to vary by discipline

Simple Example in Stata

- Data: NLSY 97
- Sample: BA degree earners
- Dependent Variable: Entry into a STEM occupation
- Independent Variable: Parent education (categorical variable of highest degree: 2-year degree or lower versus BA and Advanced Degree)

Stata Output: Logit

```
. logit stemjob  pared_ba pared_adv if sampleba==1

Iteration 0:   log likelihood =  -920.3815
Iteration 1:   log likelihood = -913.98734
Iteration 2:   log likelihood = -913.94785
Iteration 3:   log likelihood = -913.94785

Logistic regression                               Number of obs   =       2112
                                                  LR chi2(2)      =      12.87
                                                  Prob > chi2     =     0.0016
Log likelihood = -913.94785                       Pseudo R2       =     0.0070

------------------------------------------------------------------------------
     stemjob |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    pared_ba |   .4771138   .1411431     3.38   0.001     .2004784    .7537492
   pared_adv |   .3685459   .1490065     2.47   0.013     .0764986    .6605932
       _cons |  -1.920446   .0957723   -20.05   0.000    -2.108156   -1.732736
------------------------------------------------------------------------------
```

Interpretation

- Logistic Regression
  - Log odds
    - Interpretation: Among BA earners, having a parent whose highest degree is a BA degree versus a 2-yr degree or less increases the log odds of entering a STEM job by 0.477.

Interpretation

- Logistic Regression
  - Log odds
    - Interpretation: Among BA earners, having a parent whose highest degree is a BA degree versus a 2-year degree or less increases the log odds by 0.477.

  - However, we can easily transform this into odds ratios by exponentiating the coefficients: exp(0.477)=1.61
    - Interpretation: BA degree earners with a parent whose highest degree is a BA degree are 1.61 times more likely to enter into a STEM occupation than those with a parent who have a 2-year degree or less.

Stata Output: logistic

- "logistic" command outputs odds ratios instead of log odds

```
. logistic stemjob  pared_ba pared_adv if sampleba==1

Logistic regression                             Number of obs   =       2112
                                                LR chi2(2)      =      12.87
                                                Prob > chi2     =     0.0016
Log likelihood = -913.94785                     Pseudo R2       =     0.0070

------------------------------------------------------------------------------
     stemjob | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    pared_ba |   1.611417   .2274404     3.38   0.001     1.221987    2.124952
   pared_adv |   1.445631   .2154084     2.47   0.013     1.079501     1.93594
       _cons |   .1465416   .0140346   -20.05   0.000     .1214617    .1768001
------------------------------------------------------------------------------
```

Stata Output: probit

```
. probit stemjob  pared_ba pared_adv if sampleba==1

Iteration 0:   log likelihood =  -920.3815
Iteration 1:   log likelihood = -913.95526
Iteration 2:   log likelihood = -913.94785
Iteration 3:   log likelihood = -913.94785

Probit regression                                 Number of obs   =       2112
                                                  LR chi2(2)      =      12.87
                                                  Prob > chi2     =     0.0016
Log likelihood = -913.94785                       Pseudo R2       =     0.0070

------------------------------------------------------------------------------
     stemjob |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    pared_ba |   .2626883   .0779146     3.37   0.001     .1099786     .415398
   pared_adv |   .2014769   .0818166     2.46   0.014     .0411193    .3618345
       _cons |  -1.136796    .051066   -22.26   0.000    -1.236883   -1.036708
------------------------------------------------------------------------------
```

Interpretation

- Probit Regression
  - Z-scores
    - Interpretation: Among BA earners, having a parent whose highest degree is a BA degree versus a 2-year degree or less increases the z-score by 0.263.
    - Researchers often report the marginal effect, which is the change in y* for each unit change in x.

Comparison of Coefficients

Variable	Logistic Coefficient	Probit Coefficient	Ratio
Parent Ed: BA Deg	.4771	.2627	1.8
Parent Ed: Advanced Deg	.3685	.2015	1.8

Comparing Across Models

- It can be misleading to compare coefficients across models because the variance of the underlying latent variable (y^*) is not identified and can differ across models.

Some Possible Solutions to this Problem:

- Predicted Probabilities
  - Gives predicted values at substantively meaningful values of x_k
- y*-standardized coefficients
  - B_k^{sy*} gives the standard deviation increase in y* given a one unit increase in $x_{k,}$, holding all other variables constant.
- Fully standardized coefficients
  - B_k^s gives the standard deviation increase in in y*, given a one standard deviation increase in x_k, holding all other variables constant.
- Marginal effects
  - The slope of the probability curve relating x to Pr(y=1|x), holding all other variables constant

A Few Examples of Hypothesis Testing and Model Fit for Logistic Regression in Stata

- Likelihood Ratio
  - lrtest
- Wald test
  - test
- Akaike's Information Criterion (AIC)/Bayesian Information Criterion (BIC)
  - estat ic
- Or for a variety of fit statistics
  - fitstat

References

- Agresti, Alan. *An introduction to categorical data analysis. Vol. 423. Wiley-Interscience, 2007.*
- Long, J. Scott. *Regression models for categorical and limited dependent variables. Vol. 7. Sage, 1997.*
- Powers, D., and Y. Xie. "Statistical method for categorical data analysis Academic Press." *San Deigo, CA (2000).*