Binary Logit and Probit

David A. Hughes, Ph.D.

Auburn University at Montgomery

david.hughes@aum.edu

February 7, 2020

Overview

Introduction

By the time students finish this unit, they should be able to explain:

- The binary response model from a latent model perspective,
- The logic underlying logit and probit,
- How to interpret logit and probit results, and
- How to estimate and present these effects.

Review

- Suppose our dependent variable is measured such that $Y_i \in \{0, 1\}$.
- Recall that the linear probability model violates a number of desirable assumptions of OLS.
- We'd like instead to model the actual probability of observing either a "0" or "1."

A latent variable approach

- Suppose there exists an underlying measure of Y_i that is measured on a continuous scale. Call this latent variable Y_i^*. The underlying model is therefore,

$$Y_i^* = \mathbf{X_i}\beta + \epsilon_i, \tag{1}$$

where ϵ is distributed according to some normal distribution.

- We do not observe Y_i^* directly but merely its manifestations in Y_i such that:

$$\begin{aligned} Y_i &= 0 \text{ if } Y_i^* < 0 \\ Y_i &= 1 \text{ if } Y_i^* \geq 0. \end{aligned}$$

A latent approach (cont'd.)

- Taking Equation (1), we can model the probability of observing $Y_i = 1$:

$$\begin{aligned} Pr(Y_i = 1) &= Pr(Y_i^* \geq 0) \\ & Pr(\mathbf{X_i}\beta + \epsilon_i \geq 0) \\ & Pr(\epsilon_i \geq -\mathbf{X_i}\beta) \\ & Pr(\epsilon_i \leq \mathbf{X_i}\beta), \end{aligned} \tag{2}$$

  where the last inequality holds due to the symmetry of the distribution of ϵ.

- Because ϵ is assumed normally distributed, we can integrate over it to find $\hat{\beta}$.

Logit

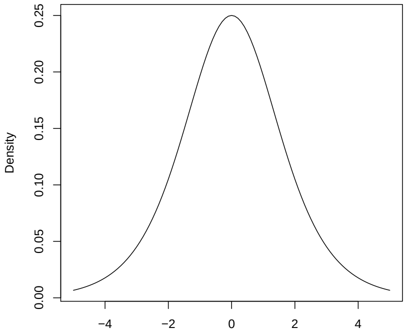
- If we assume that ϵ is distributed according to a logistic probability density function, we get the logit model:

$$Pr(\epsilon) \equiv \lambda(\epsilon) = \frac{\exp(\epsilon)}{[1+\exp(\epsilon)]^2}. \tag{3}$$

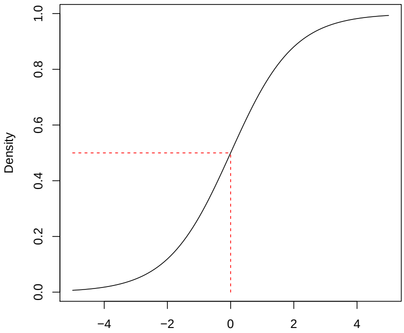
  Equation (3) gives us the pdf of the logistic distribution.
- If we want to calculate the cumulative probability that a variable distributed according to the logistic distribution is less than some value, ϵ, then we use the cumulative density function (cdf):

$$\Lambda(\epsilon) = \int_{-\infty}^{\epsilon} \lambda(\epsilon) d\epsilon = \frac{\exp(\epsilon)}{1+\exp(\epsilon)} \tag{4}$$

The logistic pdf and cdf



Logistic pdf



Logistic cdf

Specifying the logit model

- Assuming ϵ is distributed according to the standard logistic distribution, we can rewrite Equation (2):

$$Pr(Y_i = 1) \equiv \Lambda(\mathbf{X_i}\beta) = \frac{\exp(\mathbf{X_i}\beta)}{1 + \exp(\mathbf{X_i}\beta)}. \tag{5}$$

- To extract a probabilistic statement from Equation (5), we need to think about events in terms of likelihood.

Deriving the log-likelihood function for logit

$$L_i = \left(\frac{\exp(\mathbf{X_i}\beta)}{1+\exp(\mathbf{X_i}\beta)}\right)^{Y_i}\left[1-\left(\frac{\exp(\mathbf{X_i}\beta)}{1+\exp(\mathbf{X_i}\beta)}\right)\right]^{1-Y_i} \quad (6)$$

$$L = \prod_{i=1}^{N}\left(\frac{\exp(\mathbf{X_i}\beta)}{1+\exp(\mathbf{X_i}\beta)}\right)^{Y_i}\left[1-\left(\frac{\exp(\mathbf{X_i}\beta)}{1+\exp(\mathbf{X_i}\beta)}\right)\right]^{1-Y_i} \quad (7)$$

$$lnL = \sum_{i=1}^{N} Y_i ln\left(\frac{\exp(\mathbf{X_i}\beta)}{1+\exp(\mathbf{X_i}\beta)}\right) + (1-Y_i) ln\left[1-\left(\frac{\exp(\mathbf{X_i}\beta)}{1+\exp(\mathbf{X_i}\beta)}\right)\right] \quad (8)$$

- We then maximize the log-likelihood with respect to $\hat{\beta}$s to obtain our MLEs.

Probit

- If we assume that ϵ_i is distributed standard normally (i.e., $\epsilon_i \sim N(0,1)$), then we estimate a probit rather than a logit.
- Recall that the pdf of a standard normal distribution is:

$$Pr(\epsilon) \equiv \phi(\epsilon) = \frac{1}{\sqrt{2\pi}} \text{exp}\left(\frac{-\epsilon^2}{2}\right) \tag{9}$$

- The cdf for the standard normal is given by:

$$\Phi(\epsilon) = \int_{-\infty}^{\epsilon} \frac{1}{\sqrt{2\pi}} \text{exp}\left(\frac{-\epsilon^2}{2}\right) d\epsilon. \tag{10}$$

Deriving the log-likelihood function for the probit

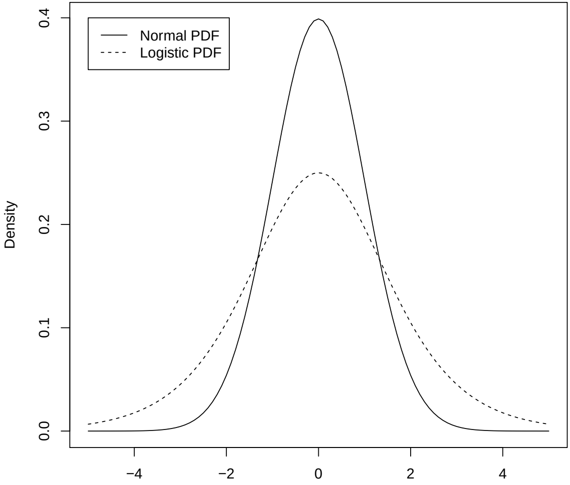
$$
\begin{aligned}
Pr(Y_i = 1) &= \Phi(\mathbf{X_i}\beta) \\
&= \int_{-\infty}^{\mathbf{X_i}\beta} \frac{1}{\sqrt{2\pi}} \text{exp}\left(\frac{-\mathbf{X_i}\beta^2}{2}\right) d\mathbf{X_i}\beta \qquad (11)
\end{aligned}
$$

- The standard normal may be a better specification for ϵ.
- But unlike the standard logistic cdf, we can't calculate the integral via a closed-form solution.
- Hence, we must use approximation methods.
- Also, we can't extract probabilities so easily as we did in logit.

Comparing logit and probit

- Each is single-peaked and symmetric.
- But logit has fatter tails than does probit.
- Logit coefficients are about 1.7 times larger than probit coefficients.
- But this turns out not to really matter.

Comparing logit and probit pdfs



Predicting election day winners

- Suppose we're interested in modeling why some candidates for office win and some lose.
- We therefore estimate the following logistic regression:

$$
\begin{aligned}
Pr(\text{Winner}_i = 1) \quad &= \quad \Lambda(\beta_0 + \beta_1 \text{Money}_i + \beta_2 \text{Incumbent}_i + \\
&\quad \beta_3 \text{Nonwhite}_i + \beta_4 \text{Female}_i),
\end{aligned}
$$

where Money_i measures a candidates campaign fundraising in millions, Incumbent_i is a dummy variable for whether the candidate is an incumbent, and Nonwhite_i and Female_i are dummy variables indicating nonwhite and female candidates, respectively.

What do we make of our logit/probit results?

```
. logit winner cm_justice_million incumbent nonwhite female

Iteration 0:   log likelihood = -444.63745
Iteration 1:   log likelihood = -282.03634
Iteration 2:   log likelihood = -274.42329
Iteration 3:   log likelihood = -274.29387
Iteration 4:   log likelihood = -274.29368
Iteration 5:   log likelihood = -274.29368

Logistic regression                      Number of obs   =        668
                                         LR chi2(4)      =     340.69
                                         Prob > chi2     =     0.0000
Log likelihood = -274.29368              Pseudo R2       =     0.3831
```

winner	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
cm_justice_million	.4805724	.1854172	2.59	0.010	.1171613	.8439834
incumbent	3.682391	.2611491	14.10	0.000	3.170549	4.194234
nonwhite	-1.072303	.3432889	-3.12	0.002	-1.745137	-.3994688
female	.5514401	.2426576	2.27	0.023	.0758401	1.02704
_cons	-1.159442	.1667965	-6.95	0.000	-1.486357	-.8325266

Interpreting probit and logit results

- "Signs and significance" (not great but better than nothing)
- Marginal effects (e.g., standardize the IVs or $\hat{\beta}$s)
- Predicted probabilities (but over what range?)

X_k's nonlinear effect on Y_i

- Recall that the estimated effect of some X_k on the DV ($\hat{\beta}_k$) is linear only with respect to the latent variable, Y_i^*.
- Hence, we cannot interpret $\hat{\beta}_k$ as a linear effect on $\hat{Y}_i$.
- The real net effect of X_k is also a function of the other variables, their coefficient estimates, and the constant:

$$\frac{\partial Pr(\hat{Y}_i = 1)}{\partial X_k} \equiv \lambda(X) = \frac{\exp(X_i\hat{\beta})}{[1 + \exp(X_i\hat{\beta})]^2}\hat{\beta}_k. \qquad (12)$$

- Unlike in OLS, then, the first derivative of the function with respect to $\hat{\beta}_k$ is non-constant.

Predicted Probabilities

- Generically, we can estimate the change in predicting a "1" across two values of X_k:

$$\Delta Pr(Y_i = 1)_{X_A \rightarrow X_B} = \frac{\exp(\mathbf{X_B}\hat{\beta})}{1 + \exp(\mathbf{X_B}\hat{\beta})} - \frac{\exp(\mathbf{X_A}\hat{\beta})}{1 + \exp(\mathbf{X_A}\hat{\beta})}, \quad (13)$$

for logits, and

$$\Delta Pr(Y_i = 1)_{X_A \rightarrow X_B} = \Phi(\mathbf{X_B}\hat{\beta}) - \Phi(\mathbf{X_A}\hat{\beta}), \quad (14)$$

for probits.

Predicted probabilities: Example

- Suppose I want to know the change in the predicted probability a candidate wins if they raise no money, versus if they raise \$1 million, versus if they raise \$2 million.
- To isolate this effect, I'll hold the other IVs equal to zero (hence, this means a non-incumbent who is a white man).

Predicted probabilities: Example (cont'd.)

$$Pr(Y_i = 1 \mid \mathbf{X_i}) \quad = \quad \Lambda(-1.16 + 0.48\text{Money}_i)$$

Someone raising no money will win with probability:

$$Pr(Y_i = 1) = \Lambda(-1.16) = \frac{\exp(-1.16)}{1 + \exp(-1.16)} = 0.24.$$

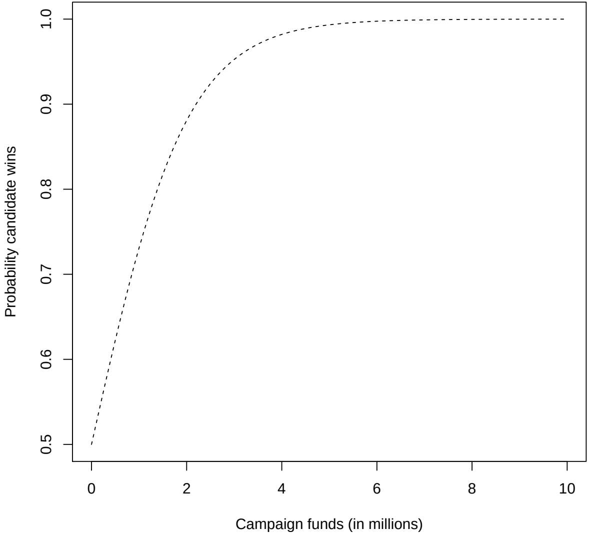
The same person who raises $1 million is predicted to win with probability:

$$Pr(Y_i = 1) = \Lambda(-1.16 + .48) = \frac{\exp(-.68)}{1 + \exp(-.68)} = 0.34.$$

And for $2 million:

$$Pr(Y_i = 1) = \Lambda(-1.16 + .96) = \frac{\exp(-.2)}{1 + \exp(-.2)} = 0.45.$$

Graphing the probability function



Estimation in Stata

- The basic syntax for estimating a logit/probit model in Stata is:
    - `logit/probit depvar [indepvars] [if] [in] [weight] [,options]`
- `if` is used to estimate for a subset of the data based on user defined conditions
- `in` is similar to `if` but defines a range of observations for which estimation is limited
- A variety of options exists with logit models (see the Stata help file for a full list), the most common relate to how the standard errors are calculated (e.g. `robust` for Huber-White robust standard errors, `cluster(groupname)` for clustered standard errors, etc.)

Interpreting $\hat{\beta}_k$ using Stata

- One of my preferred means of interpreting partial slope coefficients is using Stata's "margins" command.
- "Margins" is extremely flexible. It can provide you with marginal, standardized effects for X on Y. It can report predicted probabilities. It even comes with its own graphing environment. How awesome is that?

Adjusted predictions at the mean

```
. margins nonwhite, atmeans

Adjusted predictions                              Number of obs     =        668
Model VCE    : OIM


------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95\% Conf. Interval]
-------------+----------------------------------------------------------------
    nonwhite |
          0  |   .7386119   .0266587    27.71   0.000     .6863618    .7908621
          1  |   .4916167   .0769071     6.39   0.000     .3408815    .6423519
------------------------------------------------------------------------------
```

- This tells us that nonwhites are 24.7 percentage points less likely to win election compared to whites when holding all other variables at their means.
- But the mean of Female is 0.28. What does that mean?

Average marginal effects

```
. margins, dydx(nonwhite)

Average marginal effects                          Number of obs     =        668
Model VCE    : OIM

Expression   : Pr(winner), predict()
dy/dx w.r.t. : 1.nonwhite

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95\% Conf. Interval]
-------------+----------------------------------------------------------------
  1.nonwhite |  -.1334723   .0406578    -3.28   0.001    -.2131602   -.0537844
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

- Average marginal effect works like this: Go to $i = 1$. Treat them as if they were white. Compute the probability they win.
- Now do the same as if they were nonwhite. Take the difference in probabilities.
- Repeat for every other case in N and take the mean.

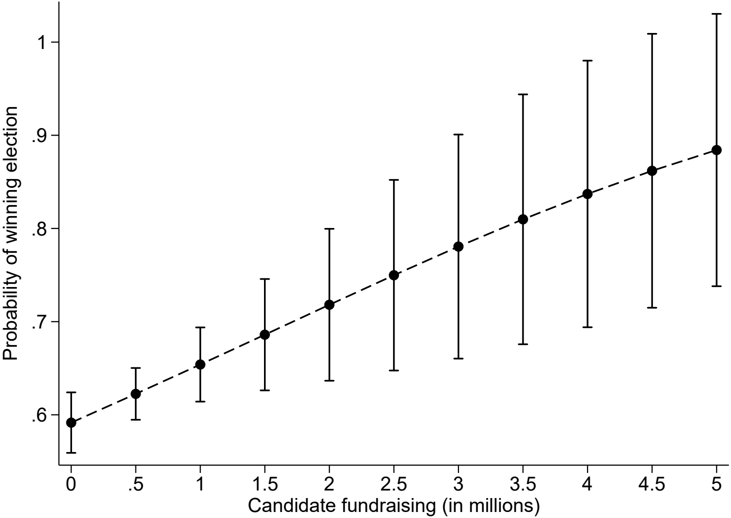
Adjusted predictions at representative values

```
. margins, at(cm_justice_million=(0, .91))

Predictive margins                              Number of obs     =        668
Model VCE    : OIM

Expression   : Pr(winner), predict()

1._at        : cm_justice~n    =           0

2._at        : cm_justice~n    =         .91

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95\% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   .5916008   .0165606    35.72   0.000     .5591427    .6240589
          2  |   .6482608   .0187586    34.56   0.000     .6114946    .6850271
------------------------------------------------------------------------------
```

Get a plot for the whole range:

```
quietly margins, at(cm_justice_million=(0(.5)5))
marginsplot
```

Adjusted predictions for continuous variable



Using margins to interpret interaction effects

- Suppose I think that when incumbents versus challengers raise vast sums of money, these events have different ramifications for the likelihood a candidate wins.
- Therefore, I reestimate the previous model, but this time, I interact Incumbent with Money ($\text{Incumbent}_i \times \text{Money}_i$).
- Unlike in OLS, in MLE, we can't just multiply variables across one another. See code below.

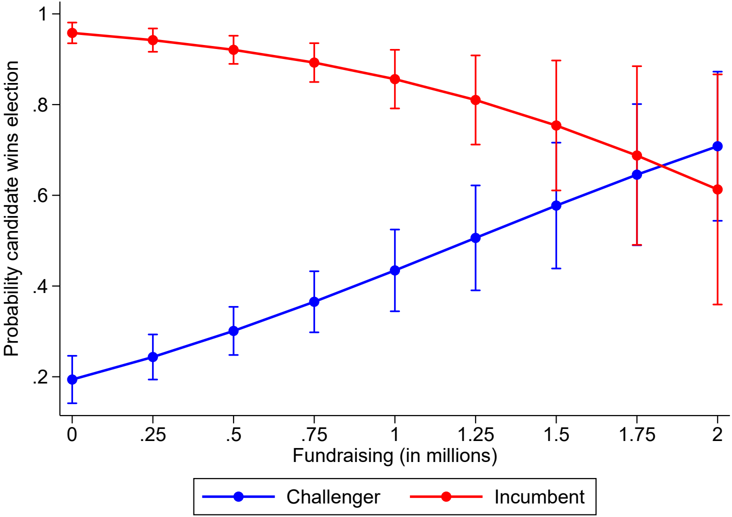
Logit with an interaction effect

```
Logistic regression                             Number of obs   =        668
                                                LR chi2(5)      =     380.06
                                                Prob > chi2     =     0.0000
Log likelihood = -254.60769                     Pseudo R2       =     0.4274
```

winner	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
cm_justice_million	1.1891	.2519415	4.72	0.000	.6953037	1.682896
1.incumbent	4.665217	.3466478	13.46	0.000	3.9858	5.344635
incumbent#c.cm_justice_million						
1	-2.553493	.4344857	-5.88	0.000	-3.405069	-1.701916
nonwhite	-.9382109	.3573678	-2.63	0.009	-1.638639	-.2377829
female	.431282	.251886	1.71	0.087	-.0624056	.9249696
_cons	-1.458338	.1846199	-7.90	0.000	-1.820187	-1.09649

```
margins incumbent, at(cm_justice_million=(0(.25)2))
marginsplot
```

Plotted interaction effect



Using Stata's "predict" commands

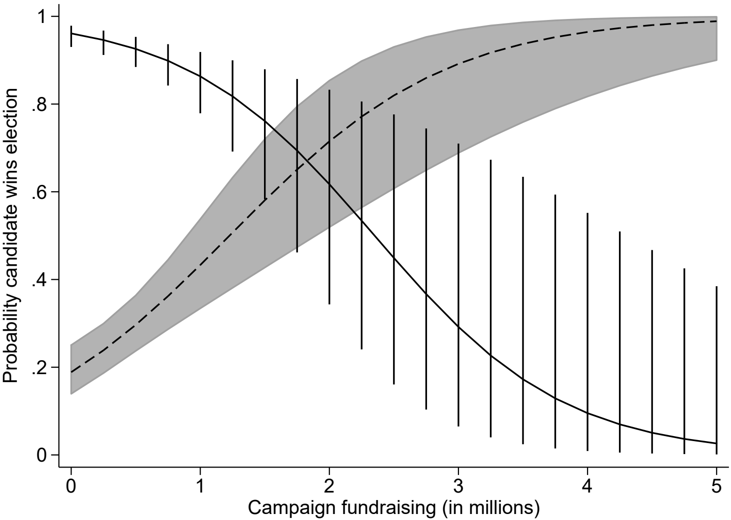
- If you want in-sample predictions, then the predict command is likely the way to go (useful for calculating proportion accurately predicted).
- Problem is, in-sample predictions can get a little ugly.

```
logit winner cm_justice_million incumbent nonwhite female

predict phat
predict sehat, stdp
predict xbeta, index
```

Out of sample predictions using Stata

- We can simulate data to create clean graphs for our variables of interest.
- Create a toy dataset that allows Money_i to vary from its min to max, and hold all other variables appropriately constant (mean for continuous variables, median else).

Out-of-sample graph



Presenting results tabularly

Table: Electoral outcomes for state judges (2002-2014)

Variable	$\hat{\beta}_k$	$\hat{\sigma}_{\hat{\beta}_k}$	$\Delta Pr(Y_i = 1)$
Money	1.19*	0.25	0.19→0.41
Incumbent	4.67*	0.35	0.19→0.96
Money × Incumbent	-2.55*	0.43	0.96→0.87
Nonwhite	-0.94*	0.36	063→0.52
Female	0.43	0.25	*n.s.*
Constant	-1.49*	0.18	—
Log-Likelihood $= -254.61$			

Notes: The dependent variable is whether a candidate won their election ("1" if yes, "0" else). Asterisks indicate statistical significance ($p < 0.05$, two-tailed). $\Delta Pr(Y_i)$ denotes the change in predicted probability (shift from min. to max for dichotomous variables, $\bar{x} - \sigma$ to $\bar{x} + \sigma$ otherwise).

Conclusion

- By now you should be prepared to estimate, interpret, and present binary response models such as logit and probit.
- Next time, we'll discuss goodness of fit and model specification.
- Remember that Homework 1 is due one week from today.