# Maximum Likelihood Estimation

Charles H. Franklin
franklin@polisci.wisc.edu

University of Wisconsin – Madison

Lecture 7

*Binary Response Models*

Last Modified: June 13, 2005



# Binary Response Models

- There are many naturally binary social outcomes:
- A citizen votes or does not.
- A cabinet forms or does not.
- A child is born or not.
- A refrigerator is bought or not.



# The Binary Response DGP

- We have already encountered a distribution for outcomes which take on only two values, the Bernoulli distribution:
- $$y_i \sim \begin{array}{c|r} y & p(y) \\ \hline 1 & \pi \\ 0 & 1-\pi \end{array}$$
- where the event occurs with probability $\pi$ and fails to occur with probability $1-\pi$.
- All our binary outcome models rest on this Bernoulli distribution of $y_i$.



# Limitations of the Bernoulli

- This could serve as a model of dichotomous choice if each event had the same chance of occurring.
- But it is not so good a model for widely variable outcomes.
- For example, it is silly to represent all voters as having the same probability of supporting Labour.
- Without some modification, the Bernoulli distribution is far too restrictive to be interesting.

## Reparameterizing the Bernoulli

- We need to let $\pi$ vary across cases. That is we need $\pi \to \pi_i$.
- This keeps the Bernoulli form but allows us to capture variation across cases in the probability.
- It also represents outcomes as *inherently* stochastic, not random only due to "error" of some kind. This is a substantively better way of thinking about behavior.



## $\pi_i =$ what?

- Reparameterizing $\pi_i$ raises the same identification issues as we saw with $\mu_i$ and $\sigma_i^2$ in regression.
- We need to write $\pi_i = f(x_i, \beta)$ in order to both reduce the number of parameters *and* to add substantive explanatory variables.
- But $\pi_i$ represents a *probability* so the reparameterization must be a probability and so must remain bounded by $(0, 1)$.
- So $\pi_i = x_i\beta$ is a bad idea since this function is unbounded and so might well fall outside the $(0, 1)$ interval.



## Functions for $\pi_i$

- Since $\pi_i$ is a probability we can take *any* probability distribution function as the basis for reparameterizing $\pi_i$.
- So long as $f(x_i, \beta)$ is a probability distribution function, it will necessarily obey the restriction that $\pi_i$ remain in the $(0, 1)$ interval.
- Because there are *many* probability distributions, this gives us many to choose from, so long as we can let $x_i\beta$ serve as the location on the distribution function.



## What could be used

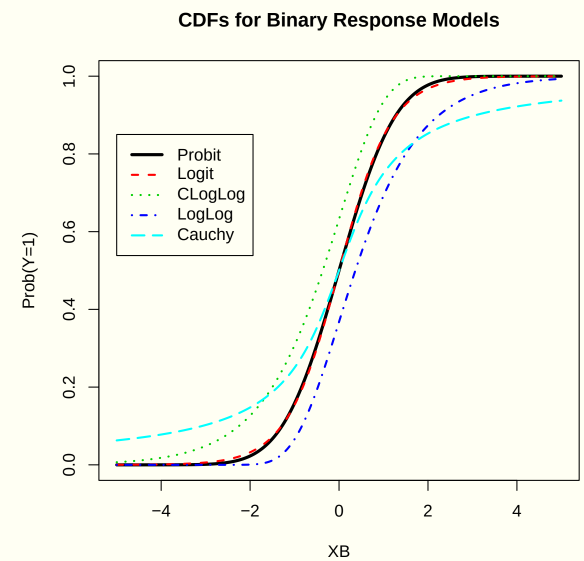
- Aside from taking a continuous, unbounded $x_i\beta$ as an argument, and returning a probability, there is really very little constraint on this function.
- While much of the literature focuses on modest (sometimes minute!) differences among symmetric distributions, somewhat less attention has been paid to alternative asymmetric forms.
- Long mentions the log-log model.
- Nagler's Scobit model offers a nice innovation by allowing the degree of asymmetry to depend on a parameter which is estimated as part of the model. (Now available in Stata.)

# Future Alternatives

- A generalization of this approach, would be to adopt a probability distribution in which a shape parameter determines the form of the distribution and use the data to estimate where marginal effects are maximized.
- A gamma distribution might be an interesting application here, since gamma can be symmetric or not, and converges to a normal under certain conditions.
- The beta distribution also offers some interesting possibilities, since it can be symmetric, skew right or left, near uniform, and even bimodal.
- Despite these potential developments, by far the most popular specifications are the normal and logistic distributions.



# What the CDFs Look Like






# The Probit Model

- The *probability density function* or pdf is the function that plots the familiar “bell shaped curve” of introductory statistics texts.
- For the normal, the pdf is

$$\phi(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

- This is the function we’ve used in constructing the joint density for the normal regression model.



# The Probit Model

- The *cumulative distribution function* or cdf, is the integral of the pdf, from $-\infty$ to a point of interest.
- This gives the probability that a realization from this distribution will be less than the point of interest. For the normal this is
- $$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) dx$$

  which gives the probability of a value less than (or equal to) $x$.

# The Probit Model

- In modelling binary outcomes, we parameterize the probability of success, $\pi_i$ as the cdf of a chosen distribution.
- So for the normal we have

$$\pi = \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) dx$$

- This integral does not have a closed form, so we usually just abbreviate it as $\Phi(x)$ and rely on numerical approximation. (See Johnson, Kotz and Balakrishnan, *Continuous Univariate Distributions*, Vol I, 2nd edition, pp. 113-121.)



# Identification in Probit Model

- The normal distribution has location parameter $\mu$ and scale parameter $\sigma^2$. There is not enough information in a binary $y$ to identify these two parameters.
- We can solve this problem by setting $\mu = 0$ and $\sigma^2 = 1$. This has the effect of rescaling the $\beta$ as $\beta/\sigma$. However, this does not change the predicted probabilities, so this is innocuous.
- The revised CDF is now:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x)^2\right) dx$$



# The Logit Model

- The (standardized) logistic distribution has pdf of

$$\lambda(x) = \frac{e^x}{[1+e^x]^2}$$

- In this case, the cdf is a very convenient closed form:

$$\begin{aligned}\Lambda(x) &= \int_{-\infty}^{x} \frac{e^x}{[1+e^x]^2} dx \\ &= \frac{e^x}{1+e^x} \\ &= \frac{1}{1+e^{-x}}\end{aligned}$$

- (Derive the last step from the previous one.)



# Binary Response Likelihood

- Let $F(x_i\beta)$ stand for either the cumulative normal or logistic.(Actually $F(x_i\beta)$ could be *any* proper cdf defined for $x_i\beta$.)
- Then the likelihood for these dichotomous choice models is:

$$\begin{aligned}L &= \prod_{i=1}^{N} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \\ &= \prod_{i=1}^{N} [F(x_i\beta)]^{y_i}[1-F(x_i\beta)]^{1-y_i}\end{aligned}$$

- Take the log of this to get the log likelihood:

$$\mathrm{lnL} = \sum_{i=1}^{N} y_i \ln F(x_i\beta) + (1-y_i)\ln(1-F(x_i\beta))$$

## Binary Response Likelihood

- Now substitute either the probit of the logit cdf for $F(x_i\beta)$.
- The probit model becomes

$$\ln L = \sum_{i=1}^{N} y_i \ln \Phi(x_i\beta) + (1 - y_i)\ln(1 - \Phi(x_i\beta))$$

- The logit model is

$$\begin{aligned} \ln L &= \sum_{i=1}^{N} y_i \ln \Lambda(x_i\beta) + (1 - y_i)\ln(1 - \Lambda(x_i\beta)) \\ &= \sum_{i=1}^{N} y_i \ln\left(\frac{1}{1 + e^{-x_i\beta}}\right) + (1 - y_i)\ln\left(1 - \frac{1}{1 + e^{-x_i\beta}}\right) \end{aligned}$$



## Binary Response Estimation

- Both of these functions are nonlinear, so no closed form solutions for $\beta$ exist, but numerical maximization is easy since both are globally concave.



## Cauchy model

- Another symmetric distribution is the Cauchy. This distribution has pdf (here $\pi \approx 3.14159$):

$$f(x) = \frac{1}{\pi(1 + x^2)}$$

- and cdf

$$F(x) = \pi^{-1}\arctan(x) + \frac{1}{2}$$

- This distribution is equivalent to a $t$-distribution with one degree of freedom. Compared to the normal, the Cauchy has very heavy tails.
- The Cauchy is also unusual because it has no moments. The expected value of the Cauchy is $\infty$, so other moments do not exist. However, the integral is perfectly well defined.



## Complementary Log-Log Model

- Two asymmetric distributions are the complementary log-log and the log-log models.
- The c-log-log CDF is $\pi_i = 1 - \exp(-\exp(x_i\beta))$
- So the log likelihood is

$$\begin{aligned} \ln L &= \sum_{i=1}^{N} y_i \ln(1 - \exp(-\exp(x_i\beta))) \\ &\quad + (1 - y_i)\ln(-\exp(-\exp(x_i\beta))) \end{aligned}$$

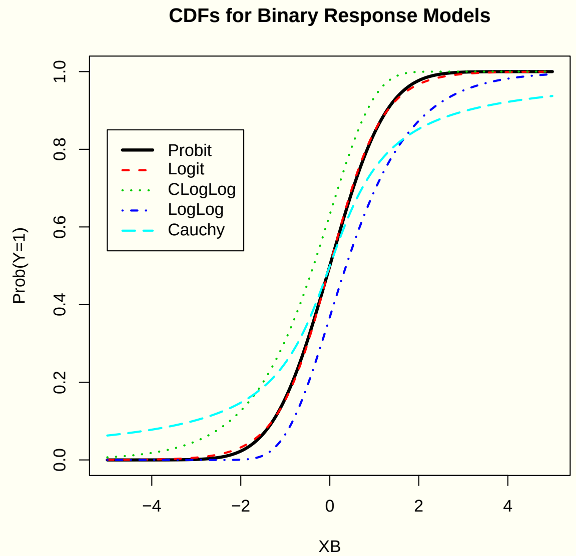
# Log-Log Model

- The CDF of the log-log model is $\exp(-\exp(-x_i\beta))$
- and log-likelihood is

$$\begin{aligned}\ln L &= \sum_{i=1}^{N} y_i \ln(\exp(-\exp(-x_i\beta))) \\ &\quad + (1-y_i)\ln(1-\exp(-\exp(-x_i\beta)))\end{aligned}$$



# What the CDFs Look Like






# How'd he do that?

```
## Create example plots of CDF for binary model
 x<-seq(-5,5,.01)
 x2<-1.7*x     # Adjust for different Logit scale
 Px<-pnorm(x)
 Lx<-1/(1+exp(-x2))
 FCLL<-1-exp(-exp(x))
 FLL<-exp(-exp(-x))
 FC<-pi^(-1)*atan(x) + .5
 postscript("c:/txt/sp/sp02/lecs/lec08/CDF1.eps",onefile=FALSE, hor=FALSE, wid=6,he=6)
 plot(Px ~x,type="l",lty=1,col=1,lwd=3,
     main="CDFs for Binary Response Models",
     xlab="XB",
     ylab="Prob(Y=1)")
 lines(Lx~x,type="l",lty=2,col=2,lwd=2)
 lines(FCLL~x,type="l",lty=3,col=3,lwd=2)
 lines(FLL~x,type="l", lty=4,col=4,lwd=2)
 lines(FC~x,type="l",lty=5,col=5,lwd=2)
 leg.txt=c("Probit","Logit","CLogLog","LogLog","Cauchy")
 legend(-5,.85,leg.txt,lty=c(1,2,3,4,5),
     col=c(1,2,3,4,5),lwd=c(3,2,2,2,2))
 dev.off()
```



# Probit vs. Logit

- Very little important difference between these two parameterizations.
- The simplicity of the logit model provides some modest edge.
- When extensions such as heteroskedasticity are considered, the normal cdf of the probit becomes more tractable than the logit.
- But when extended to multiple outcomes the logit is more tractable than probit!

# Example: AFLCIO PAC Contributions

- 1992 contributions to 347 incumbent House members.
- `give` is 1 if a contribution was made, 0 otherwise.
- Independent variables are years in office (`senior`), vote won in 1990 (`vote90`) and ideological distance of member from the AFLCIO (`distance`).
- I fit both a probit and a logit model.



# Comparative Results

| | Probit | | | Logit | | |
|---|---|---|---|---|---|---|
| Variable | Coefficient | SE | Z | Coefficient | SE | Z |
| Constant | 2.64868 | 0.47895 | 5.530 | 4.64165 | 0.84098 | 5.519 |
| Seniority | -0.03138 | 0.01083 | -2.898 | -0.05306 | 0.01839 | -2.886 |
| Vote 1990 | -0.01713 | 0.00631 | -2.715 | -0.02997 | 0.01078 | -2.779 |
| Distance | -1.63813 | 0.18312 | -8.945 | -3.00226 | 0.37883 | -7.925 |

- Comparison of Probit and Logit Estimates for AFLCIO PAC contributions model.
- While the coefficients differ in size due to different scalings of the normal and logistic distributions, the substantive conclusions (and the predicted probabilities) are very similar.