POLI 8501
Binary Logit & Probit, I

The LPM, Logit, and Probit... Consider the case of a binary response ("dependent") variable...

- Happens a lot (mainly because anything can be dichotomized).
- Some variables are intrinsically binary; others are binary "versions" of some other variable...
  - Voting for a Republican or a Democrat,
  - Going to war or not,
  - Whether a country is a democracy or not (sort of...),

etc.

How do we deal with such response variables?

The Linear Probability Model

Think about the general case of linear regression, in which we model the expected value of Y as a linear function of the independent variable(s) $\boldsymbol{X}$:

$$E(Y) = \mathbf{X}\boldsymbol{\beta} \tag{1}$$

What is the expected value of Y in the binary case?

$$\begin{aligned} E(Y) &= 1[\Pr(Y=1)] + 0[\Pr(Y=0)] \\ &= \Pr(Y=1) \end{aligned} \tag{2}$$

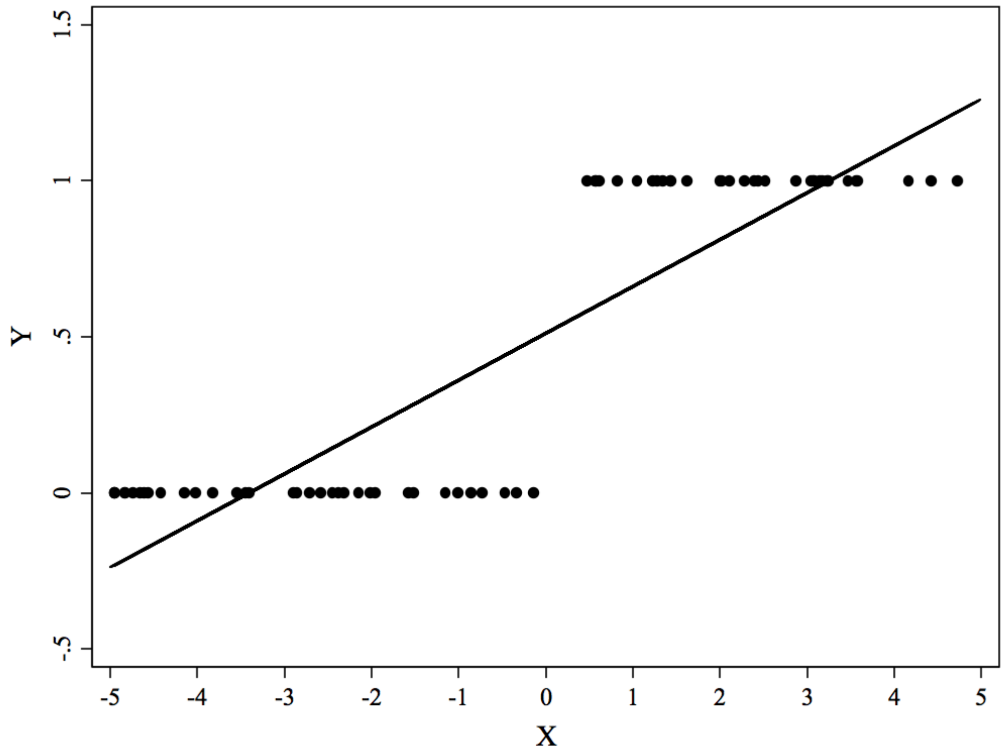
So we model:

$$\Pr(Y_i = 1) = \mathbf{X}_{\boldsymbol{i}}\boldsymbol{\beta} \tag{3}$$

That is, we just do a linear OLS regression of Y on X. This is the linear probability model (LPM):

- It amounts to fitting a regular OLS regression to a binary response variable.
- The result (in one continuous independent variable) looks like Figure 1.

Figure 1: A Linear Probability Model



Note: There's nothing whatsoever preventing you from running a LPM:

- The model will estimate properly, and give you a set of estimated $\hat{\boldsymbol{\beta}}$s.
- You can interpret those coefficients in the standard way: as the change in $\Pr(Y = 1)$ associated with a one-unit change in $\mathbf{X}$.

There are, however, a large number of **problems** with the LPM...

1. For example, what's the variance term? Since Y is binomial, we can write its variance as

$$\begin{aligned} \text{Var}(Y) &= \text{E}(Y)[1 - \text{E}(Y)] \\ &= \mathbf{X}_i\boldsymbol{\beta}(1 - \mathbf{X}_i\boldsymbol{\beta}) \end{aligned} \tag{4}$$

   - So, the variance of Y depends on the values of X and $\boldsymbol{\beta}$...

- That is, Y is *by construction* heteroscedastic.
- This spells bad juju...

2. Relatedly, the errors u_i can only take on two values: either $1 - \mathbf{X}_i\hat{\boldsymbol{\beta}}$ or $-\mathbf{X}_i\hat{\boldsymbol{\beta}}$. (why?).

   - So, they can *never* be normally distributed...
   - This can have bad effects on hypothesis tests.

3. The model can also lead to *nonsensical predictions* (i.e., probabilities that are < 0 or > 1).

4. Finally, there's the matter of *functional form.*

   - We'd usually expect that variables which impact probabilities would exhibit diminishing returns: as their value increased, the impact on Y would decrease.
   - The LPM doesn't allow this to happen; $\frac{\partial \text{E}(Y)}{\partial \boldsymbol{\beta}}$ is constant for all values of $\mathbf{X}$ and Y.

Note that (with the possible exception of the last one) *none of these problems go to bias*; OLS point estimates $\hat{\boldsymbol{\beta}}$ remain unbiased estimates of the true parameter values $\boldsymbol{\beta}$.

All the same, if you try to do an LPM for a journal submission these days, you'll get creamed...

Instead, we'll derive the models for *logit* and *probit.* In fact, we'll do so in several ways:

- First, using a latent variable conceptualization,
- Second, as a model that is log-linear in the odds of the event of interest, and
- Third, and perhaps most usefully, as the product of a model of individual choice based on utility maximization.

Logit and Probit: A Latent Variable Approach

The latent variable approach treats dichotomous dependent variables as essentially a problem of measurement.

- That is, there exists a continuous underlying variable; we just haven't measured it.
- Instead, we have a dichotomous indicator of that underlying (latent) variable.

Call this latent variable Y^*.

The underlying model is then:

$$Y_i^* = \mathbf{X}_i\boldsymbol{\beta} + u_i \tag{5}$$

This model has the usual OLS-type assumptions; in particular, that u_i is distributed according to some symmetrical distribution (e.g., Normal – more on this later).

However, we observe only the following realization of Y^*:

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

So, we can write:

$$\begin{aligned} \Pr(Y_i = 1) &= \Pr(Y_i^* \geq 0) \\ &= \Pr(\mathbf{X}_i\boldsymbol{\beta} + u_i \geq 0) \\ &= \Pr(u_i \geq -\mathbf{X}_i\beta) \\ &= \Pr(u_i \leq \mathbf{X}_i\beta) \end{aligned} \tag{6}$$

where the last equality holds because of the symmetry of the distribution of the u_is.

In other words, $Y = 1$ if the "random part" is less than (or equal to) the "systematic part."

BUT remember that u_i and $\hat{\boldsymbol{\beta}}$ are themselves *random variables*...

Now, how do we figure out this probability?

- If we assumed that u followed some distribution, we could integrate over that distribution to get an idea of the probability that u_i fell above some point (e.g. $\mathbf{X}_i\boldsymbol{\beta}$).
- This is exactly what we do...

Logit

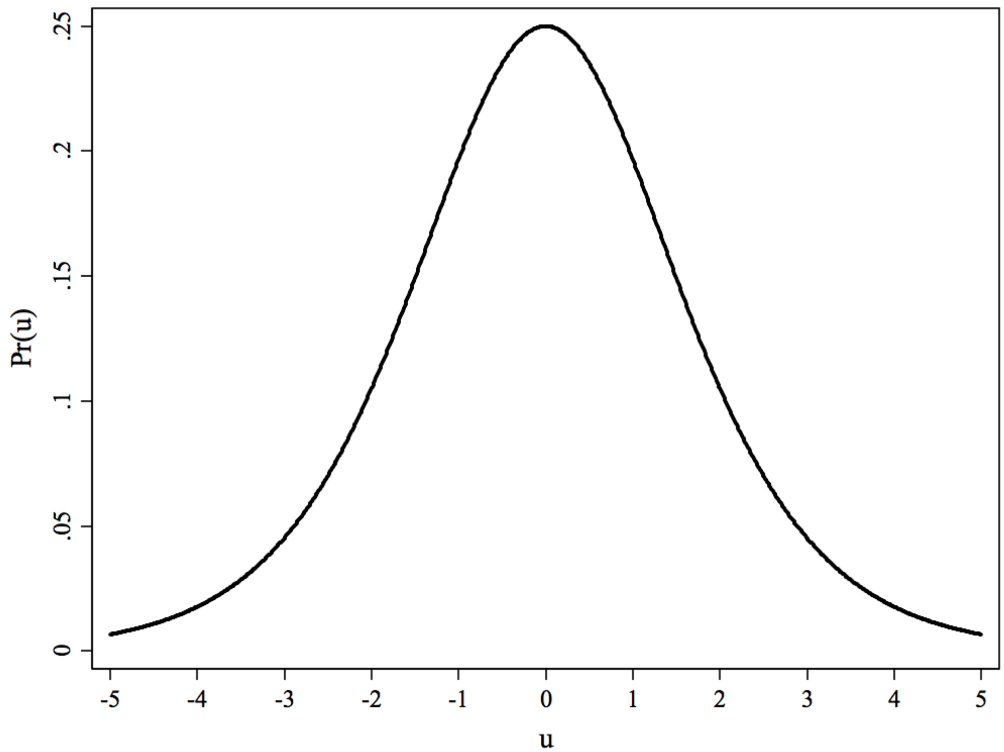
If we assume that the u_is follow a *standard logistic distribution*, we get a *logit* model...

- The standard logistic probability density is:

$$\Pr(u) \equiv \lambda(u) = \frac{\exp(u)}{[1 + \exp(u)]^2} \tag{7}$$

That's the *probability density function* (*PDF*), the term for the probability that u takes on some specific value. The standard logistic PDF is "bell-shaped," and centered around zero

Figure 2: A Logit PDF Curve



(which value has $\Pr(u) = 0.25$). It looks like the thing in Figure 2.

If we want to know the cumulative probability that a standard-logistic variable is less than or equal to some value (say, u), we consider the *cumulative distribution function* of the logit, which is:

$$\Lambda(u) = \int \lambda(u)du = \frac{\exp(u)}{1 + \exp(u)} \tag{8}$$

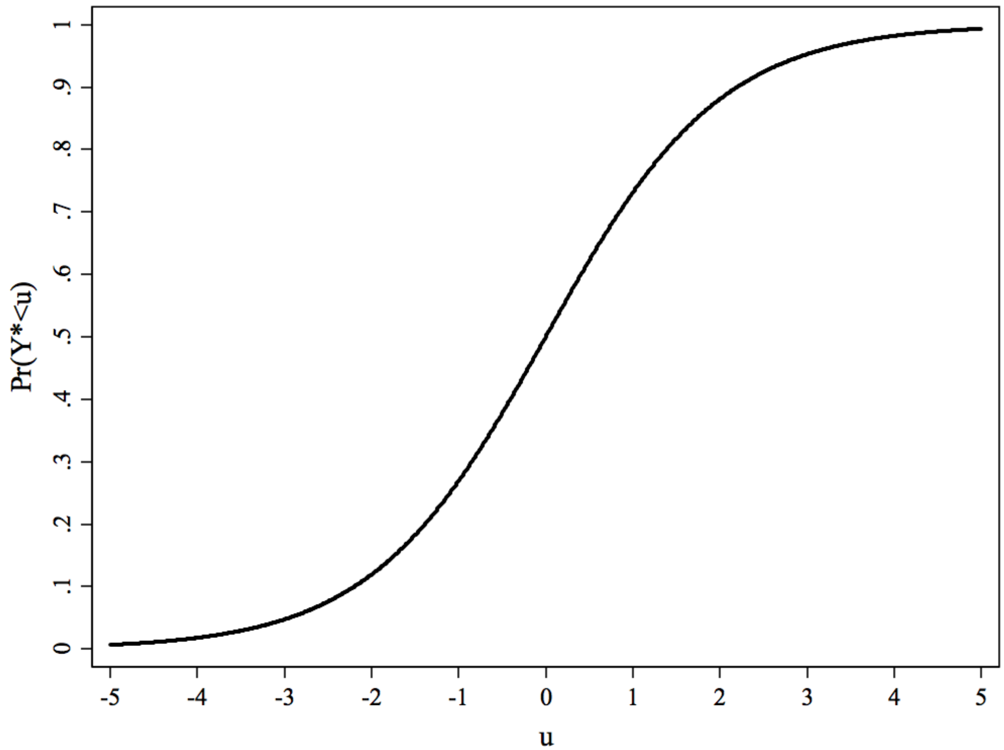
This function is the probability that a variable distributed as standard logistic will take on a value less than or equal to some value u.

As an example, consider some values for u:

If	$u = 0$	then	$\Lambda(u) = 0.5$
	$u = 1$	then	$\Lambda(u) = 0.73$
	$u = 2$	then	$\Lambda(u) = 0.88$
	$u = -1$	then	$\Lambda(u) = 0.27$
	$u = -2$	then	$\Lambda(u) = 0.12$

Generally, this looks like an S-shape:

Figure 3: A Logit CDF Curve



Characteristics of the Standard Logistic Distribution

- It is symmetrical around zero [i.e., $\lambda(u) = 1 - \lambda(-u)$ and $\Lambda(u) = 1 - \Lambda(-u)$].
- The PDF has a variance equal to $\frac{\pi^2}{3} \approx 3.29$.

Motivating the Logit Model

Once we've assumed that the error term u_i follows a standard logistic distribution, we can start to say something about the probability that $Y = 1$...

Remember that:

$$\begin{aligned} \Pr(Y_i = 1) &= \Pr(Y_i^* > 0) \\ &= \Pr(u_i \leq \mathbf{X}_i\boldsymbol{\beta}) \end{aligned}$$

We can further write this now, as:

$$\Pr(Y_i = 1) \equiv \Lambda(\mathbf{X}_i\boldsymbol{\beta}) = \frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i\boldsymbol{\beta})} \tag{9}$$

This is the basic form of the probability for the logit model. To get a probability statement for every observation i in our data, we want to think of the probability of getting a zero (one) given the values of the covariates and the parameters. That is, the likelihood for a given observation i is:

$$L_i = \left(\frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i\boldsymbol{\beta})}\right)^{Y_i} \left[1 - \left(\frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i\boldsymbol{\beta})}\right)\right]^{1-Y_i} \tag{10}$$

That is, observations with $Y = 1$ contribute $\Pr(Y_i = 1|\mathbf{X}_i)$ to the likelihood, while those for which $Y = 0$ contribute $\Pr(Y_i = 0|\mathbf{X}_i)$. Assuming that the observations are conditionally independent, we can take the product over the N observations in our data to get the overall likelihood:

$$L = \prod_{i=1}^{N} \left(\frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i\boldsymbol{\beta})}\right)^{Y_i} \left[1 - \left(\frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i\boldsymbol{\beta})}\right)\right]^{1-Y_i} \tag{11}$$

Taking the natural logarithm of this yields:

$$\ln L = \sum_{i=1}^{N} Y_i \ln\left(\frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i\boldsymbol{\beta})}\right) + (1 - Y_i)\ln\left[1 - \left(\frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i\boldsymbol{\beta})}\right)\right] \tag{12}$$

We can then maximize this log-likelihood with respect to the $\hat{\boldsymbol{\beta}}$s to obtain our MLEs, in the manner in which we discussed last week.

Um, what's a "logit," anyway?...

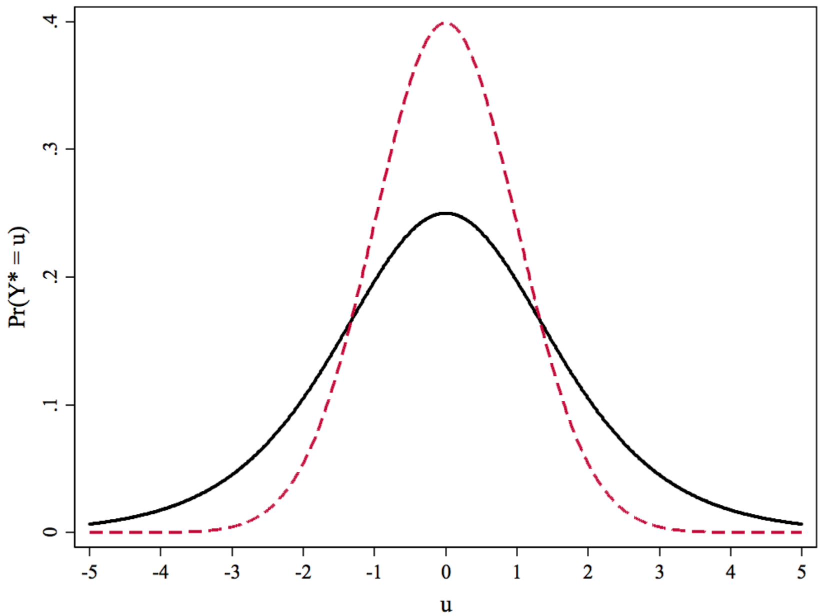
Now, *why would we ever assume that the errors follow a standard logistic distribution?*

That's a good question. In fact, it turns out that the PDF for the standard logistic looks a lot like that for the standard normal:

- Both are symmetric around zero, "bell-shaped," etc.
- The main difference is that the standard logistic has fatter "tails":

That leads us to...

Figure 4: Logistic (Smooth) and Normal (Dashed) PDFs



The Normal Distribution and Probit

We've talked about the normal distribution before...

The *standard normal* is just a normal distribution with mean zero and $\sigma^2 = 1$. It looks like:

$$\Pr(u) \equiv \phi(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \tag{13}$$

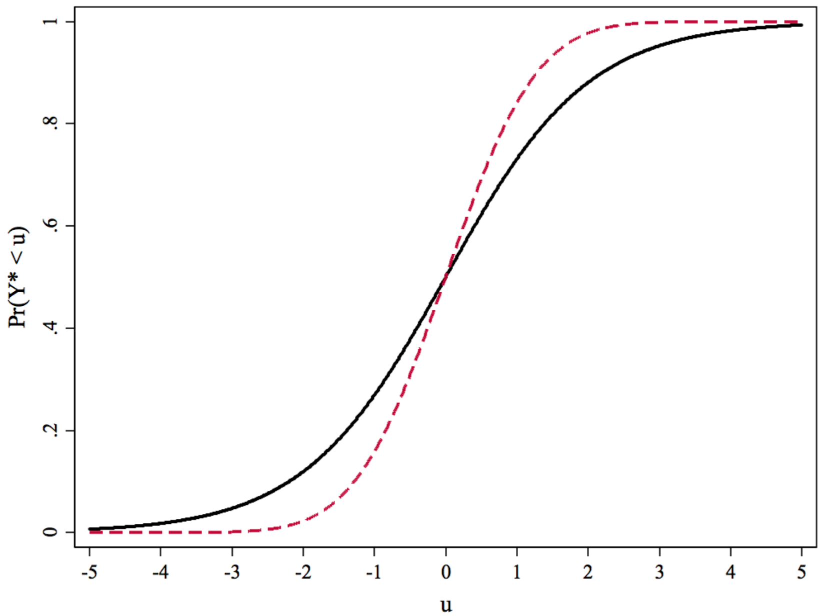
Again, this is the *density* (PDF); the probability that a standard normal random variable takes on a specific value u. The cumulative distribution function (*cdf*) is:

$$\Phi(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du \tag{14}$$

If we assume that the errors in the underlying latent equation are distributed normally, then we get the *probit* model...

- Stands for "probability logit," or something like that.
- Probably would be better called the "normit."

Figure 5: Logistic (Smooth) and Normal (Dashed) CDFs



The probability setup for the probit thus looks like:

$$\begin{aligned} \Pr(Y_i = 1) &= \Phi(\mathbf{X}_i\boldsymbol{\beta}) \qquad (15) \\ &= \int_{-\infty}^{\mathbf{X}_i\boldsymbol{\beta}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\mathbf{X}_i\boldsymbol{\beta})^2}{2}\right) d\mathbf{X}_i\boldsymbol{\beta} \end{aligned}$$

and the corresponding (log)likelihood is:

$$\ln L = \sum_{i=1}^{N} Y_i \ln\Phi(\mathbf{X}_i\boldsymbol{\beta}) + (1 - Y_i) \ln\Phi(\mathbf{X}_i\boldsymbol{\beta}) \qquad (16)$$

This formulation has its good points and its bad points...

- Some would argue that the assumption of normal errors is better justified than that of (standard) logistic errors (because of the central limit theorem, etc.).
- OTOH, the CDF of the standard normal has no closed form solution.
    - That is, one can't figure out the integral through algebra/calculus.
    - Instead, we have to approximate it numerically (hence, why we have z-tables).
    - We can approximate it very accurately, however, BUT
    - We also can't simply figure out probit probabilities using a garden-variety hand calculator.

Another Way of Thinking About Logit: Odds

Suppose that I were to ask you the *odds* of getting two heads on two consecutive flips of a fair coin. The *probability* of this happening is 0.25; we could express the odds in a number of ways:

- We'd be most likely to say that the odds are "one in four," or
- That the odds are "three to one against" getting the two consecutive heads.

What we really mean is that the odds of a given event are:

$$\Omega \equiv \text{Odds(Event)} = \frac{\Pr(\text{Event})}{1 - \Pr(\text{Event})}.$$

Since probabilities are necessarily bounded between zero and one, odds are also lower-bounded at zero, though odds have no upper bound. They are also badly skewed (why?). For these two reasons, it is often useful to think about the (natural) log of the odds of some event Z:

$$\ln[\Omega(Z)] = \ln\left[\frac{\Pr(Z)}{1 - \Pr(Z)}\right] \tag{17}$$

The log-odds has a particularly nice functional form, as illustrated in Figure 6. In particular, we can think of allowing the log-odds of the occurrence of some binary event of interest to be linearly related to a set of covariates:

$$\ln[\Omega(Z_i)] = \mathbf{X}_i\boldsymbol{\beta} \tag{18}$$

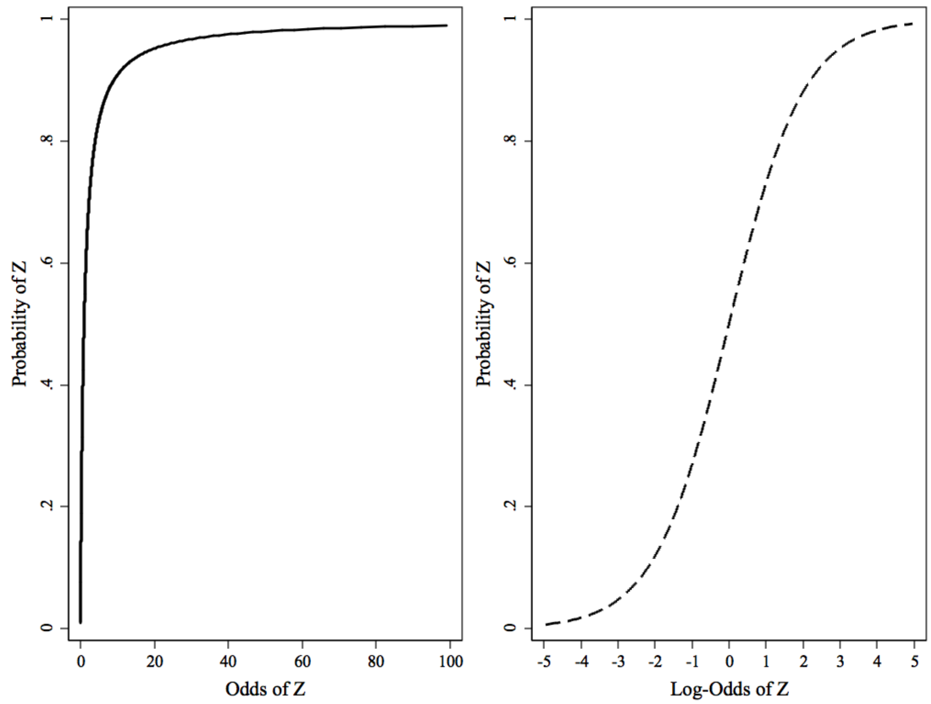
Substituting this into (17) and doing a bit of algebra, we find that:

$$\Pr(Z_i) = \frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i\boldsymbol{\beta})} \tag{19}$$

That is, positing a linear relationship between the log-odds of the (binary) event of interest and the covariates yields a standard binary logit model. There are (at least) two useful aspects to this way of thinking about logit models this way:

1. It answers the question, "Why should we think that anything is ever distributed according to a standard logistic distribution?" (The answer, of course, is "We shouldn't"). The log-odds formulation of the logit model gives us a separate, independent, and intuitively-appealing rationale for using logit.
2. As we'll see next week, this formulation also has some nice properties when it comes to interpreting the results of our models.

Figure 6: Odds and Log-Odds



Yet Another View: The Random Utility Model

Logit and probit models are often used to model dichotomous choices, and the motivation for doing so can be thought of as having its roots in microeconomic theories of choice. As a simple example, consider an individual i with some level of utility over two possible choices, the status quo (SQ) and an alternative (A).

- Assume further that i chooses the alternative which maximizes his or her expected utility. So:

$$
\begin{aligned}
Y_i &= A \text{ if } \text{E}[\text{U}_i(A)] \geq \text{E}[\text{U}_i(SQ)] \\
&= SQ \text{ if } \text{E}[\text{U}_i(SQ)] > \text{E}[\text{U}_i(A)]
\end{aligned}
$$

- The utility of each choice is determined by factors relating to i and/or to the alternative A (relative to the status quo).
  - Call these factors a vector of covariates $\mathbf{X}_{iA}$.
  - These factors are weighted by their relative importance $\boldsymbol{\beta}$.

- If the decision-making takes place under uncertainty about the relative value of the alternative as compared to the status quo, we can introduce a random (stochastic) element into the equation...

$$\mathrm{E}[\mathrm{U}_i(A)] = \mathbf{X}_{iA}\boldsymbol{\beta} + u_{iA} \tag{20}$$

So:

$$\begin{aligned} \Pr(Y = A) &= \Pr\{\mathrm{E}[\mathrm{U}_i(A)] \geq \mathrm{E}[\mathrm{U}_i(SQ)]\} \\ &= \Pr\{(\mathbf{X}_{iA}\boldsymbol{\beta} + u_{iA}) \geq \mathrm{E}[\mathrm{U}_i(SQ)]\} \end{aligned} \tag{21}$$

Now, if we set i's valuation of the status quo to some arbitrary number (say, zero), we get:

$$\begin{aligned} \Pr(Y = A) &= \Pr\{(\mathbf{X}_{iA}\boldsymbol{\beta} + u_{iA}) \geq 0\} \\ &= \Pr\{u_{iA} \geq -\mathbf{X}_{iA}\boldsymbol{\beta}\} \end{aligned}$$

which is effectively identical to (6).

Q: Why bother, if it gives us the same thing?

One reason is that we can use the *theory* to specify a distribution for the u_{iA}s...

- E.g., it might be reasonable to assume that the us are mean-zero normally distributed; this is why applied economists tend to use probits rather than logits.
- Also, we might believe that, if the stochastic element regarding information about alternative A is based on some other factors (e.g. the amount of information you have about the alternative); you can then explicitly model this...
- Also, this approach opens up the possibility of choosing among more than two alternatives; something we'll talk about much more later in the course.

A Short Digression: Other Models for Binary Responses

While we don't hear about them often, there are a number of other model forms for binary responses. I'll just mention a couple of them here; we'll return to this fact when we discuss generalized linear models at the end of the term.

The Complementary Log-Log Model

As an alternative to the logit and probit CDFs, consider a model of the form:

$$\Pr(Y_i = 1) = 1 - \exp[-\exp(\mathbf{X}_i\boldsymbol{\beta})] \tag{22}$$

or, alternatively:

$$\ln\{-\ln[1 - \Pr(Y_i = 1)]\} = \mathbf{X}_i\boldsymbol{\beta} \tag{23}$$

The is the *complementary log-log* model, sometimes abbreviated as *cloglog* (or *c-log-log*); it's called that because of the form of (23).

A key difference between the cloglog model and logit/probit is that the former is not symmetrical around 0.5. That is, for a given value of u, it is not the case that $\Pr(Y_i = 1|X_i = u) = 1 - \Pr(Y_i = 1|X_i = -u)$. Instead, the CDF is asymmetrical, with the function approaching 0 slowly, but approaching 1 quickly. An alternative, closely related model takes the form:

$$\Pr(Y_i = 1) = \exp[-\exp(\mathbf{X}_i\boldsymbol{\beta})] \tag{24}$$

or

$$\ln\{-\ln[\Pr(Y_i = 1)]\} = \mathbf{X}_i\boldsymbol{\beta} \tag{25}$$

This is often known as the *log-log* model; it isn't "complementary" because it lacks the "1-" part of the model in (22). The log-log model "inverts" the cloglog function; that means that:

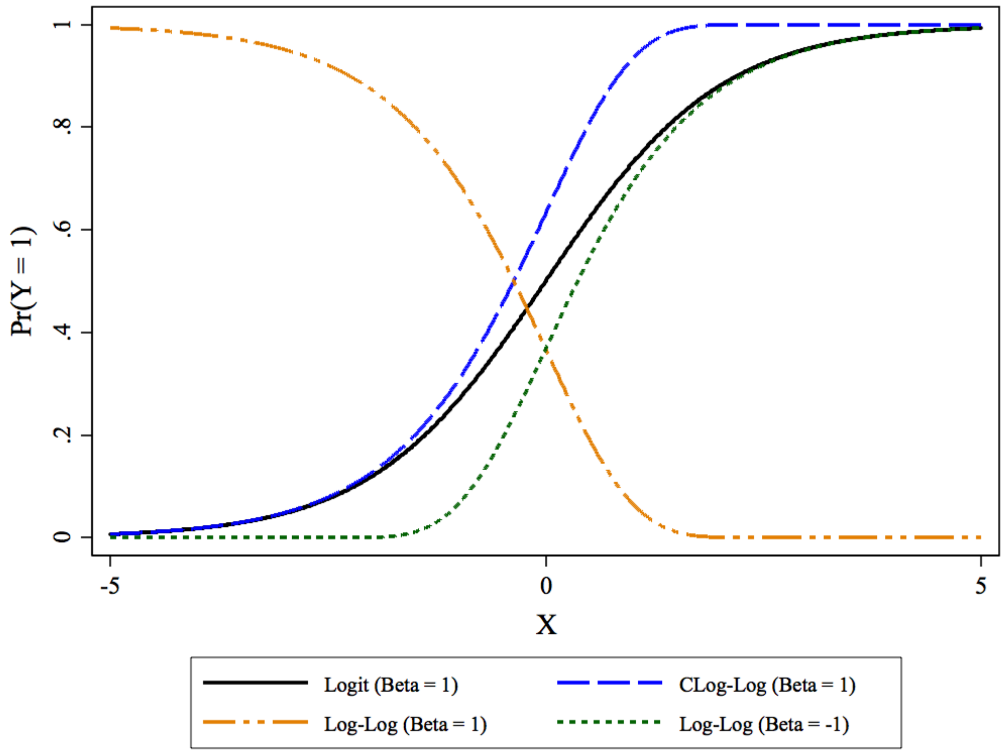
- the function approaches 1 slowly, but approaches 0 quickly. In addition,
- If $\beta > 0$, then larger values of X correspond to lower probabilities of $Y = 1$; likewise, if $\beta < 0$, then larger values of X correspond to higher probabilities of $Y = 1$. In other words, the signs of the βs are "flipped" from what we'd normally expect them to be.

Some examples of log-log and cloglog curves are presented in Figure 7. We'll actually come back to the cloglog model a bit later on, when we discuss models for event counts.

Scobit

Scobit is short for "skewed logit," and is due to Nagler (1994). Nagler starts by noting that, for both probit and logit, $\max\left(\frac{\partial \Pr(Y=1)}{\partial X}\right)$ occurs at 0.5. That is, the greatest effect of X on Y is constrained to be for observations which have a 50-50 probability of $Y = 1$. Note that this is not the case for the cloglog model, since its CDF is not symmetrical around 0.5, but it is the case that the cloglog model has a fixed point at which $\max\left(\frac{\partial \Pr(Y=1)}{\partial X}\right)$ will always be greatest.

Figure 7: Logit, Log-Log, and Complementary Log-Log CDFs



Nagler's idea, then, is to generalize logit to allow for an asymmetrical CDF, but to do so in a general way. The solution was to use a little-known distribution known as the Burr-10; the associated CDF for that distribution is:

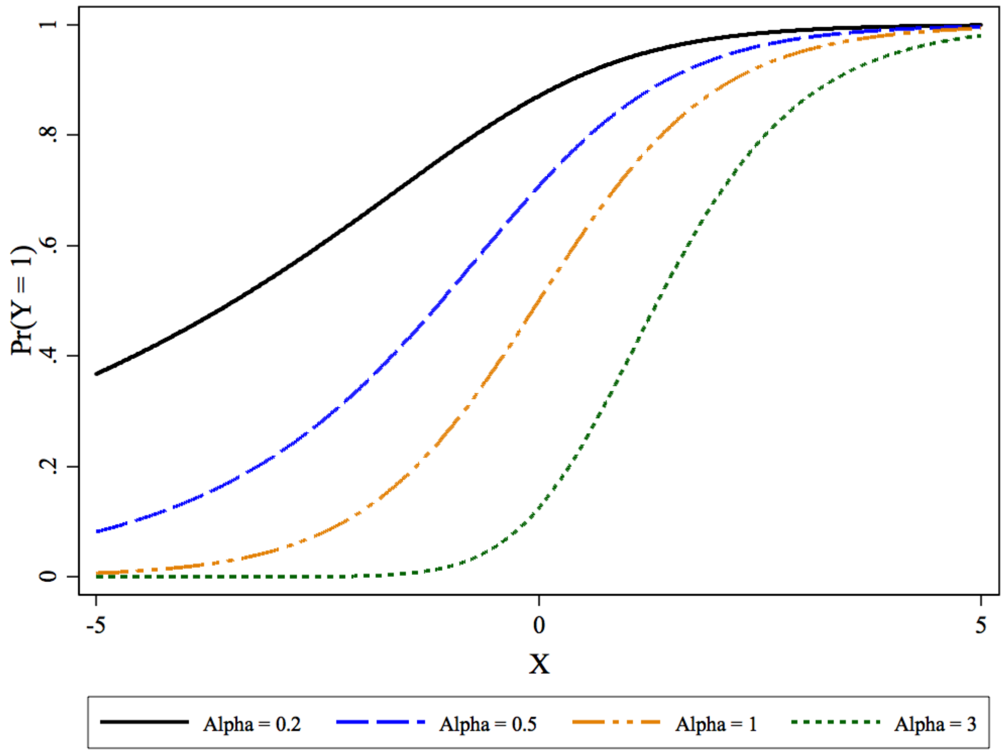
$$\Pr(Y_i = 1) = \frac{1}{[1 + \exp(-\mathbf{X}_i \boldsymbol{\beta})]^{\alpha}} \tag{26}$$

where $\alpha > 0$. This latter "shape" parameter is then flexibly estimated along with the other parameters of the model (i.e., the $\boldsymbol{\beta}$s).

Note that the scobit model "nests" the logit model as a special case when $\alpha = 1$:

$$\begin{aligned} \frac{1}{[1 + \exp(-\mathbf{X}_i \boldsymbol{\beta})]^1} &= \frac{1}{1 + \exp(-\mathbf{X}_i \boldsymbol{\beta})} \\ &= \frac{\exp(\mathbf{X}_i \boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i \boldsymbol{\beta})} \end{aligned}$$

Figure 8: Scobit CDFs, with $\beta = 1$ and Varying αs



As illustrated in Figure 8, the effect of changing α is to change the "shape" of the CDF curve. More specifically, different values of α yield CDFs for which $\max\left(\frac{\partial \Pr(Y=1)}{\partial X}\right)$ occurs at different values of $\Pr(Y = 1)$.

The scobit models have seen limited use in political science (and some other social sciences). It is, in principle, a more flexible model than plain-vanilla logit, and there is a Stata package (creatively titled `scobit`) that will estimate the model. However, people have also reported some computational difficulties with the model, since it can sometimes be the case that changes in the overall likelihood can be equally caused by changes in α or by changing one or more of the βs.

A Few General Things To Consider About Logit And Probit Models

Identification

In order to be able to get unique estimates for the parameters, we have to make a few assumptions about the model...

Why?

- In the regression context, its clear that the size of the coefficients depend on the variance of the dependent variable...
- Consider if the underlying variable isn't latent...
  - Two continuous variables: Y^* = personal income and Z^* = shoe size.
  - Dichotomize them so that observations falling above the median get a value of one, and the rest get zeros...
  - Both dichotomous variables will have $\mu = 0.5$, and identical variances ($\sigma^2 = 0.25^2$).
  - Its impossible to "get back to" the mean or the variance of either variable on the basis of the 0/1 indicators alone.
- Important: **dichotomizing continuous variables throws away information, and replaces data with assumptions** (cf. Royston et al. 2006).

When the underlying variable really is latent, then because we don't know its mean and variance we have to make assumptions about them.

- This is necessary to "fix" the model in the probability space...
- Assumptions are both arbitrary, and necessary (typically a bad combination), BUT
- They turn out not to be all that critical...

Three Assumptions

1. That the "threshold" point for going from $Y = 0$ to $Y = 1$ is $Y^* > 0$.
   - In this context, this is not a big deal, provided we include an intercept in the model.
   - Will become (slightly) more important when we discuss ordinal outcomes...
2. The conditional mean of the errors is zero; i.e. $\mathrm{E}(u_i|\mathbf{X}, \boldsymbol{\beta}) = 0$ (again, this is not a problem provided one has an intercept in the model).
3. Assume that the variances are either $\frac{\pi^2}{3}$ or one (for logit and probit, respectively).

Q: Are these assumptions important?

A: As it turns out, not really...

Logit and probit are both examples of *estimable functions*.

- This means that, conditional on the identifying assumptions about the first two moments of Y^*, we can estimate the $\boldsymbol{\beta}$s.
- Put another way: the estimates of $\hat{\boldsymbol{\beta}}$ we get depend both on the data, and on the assumptions we make about the mean and variance of Y^*.
- That is, because we don't have information on $\sigma^2_{Y^*}$, we can only estimate $\frac{\hat{\boldsymbol{\beta}}}{\sigma}$ where we assume a value for σ, but not $\hat{\boldsymbol{\beta}}$ alone.
- We typically set $\sigma^2 = 1$ for probit, since it gives us "easy" estimates of $\frac{\hat{\boldsymbol{\beta}}}{\sigma}$.

However,

The probabilities that result do not depend on the identifying assumptions.

Think of it this way...

- Suppose we have a regular probit model, where the latent variance is assumed to be one, and that we have data which are a dichotomized version of some large, continuous variable Y^*.
- If we assume that the variance of the underlying variable (and therefore, the variance term in the model) is, say, 10 instead of 1.0, we'll get a "dependent variable" with a lot more variability.
- And just as in OLS, we'll also get bigger coefficients, since the independent variables will be associated with more variability in Y^*, ***but***
- When we go to convert the latent variable back into (say) predicted probabilities of $Y_i = 1$, we'll get exactly the same predictions as if we'd assumed that $\sigma^2 = 1$.

Comparing Logit And Probit

Note at the outset that, for all intents and purposes, one is as good as the other. That said, there are a few things to keep in mind when using them.

- While the "scale" of the dependent variable is the same, the sizes of the coefficients estimated by logit and probit models will not be identical.
- Because the standard logistic distribution has a larger variance, it will also yield larger estimated $\hat{\beta}$s (and correspondingly larger standard error estimates).
- As a result, logit estimates are usually roughly 1.8 times the size of probit ones (which is $\approx \sqrt{3.29}$ – recall that the variance of the standard normal is zero, while that for the standard logistic is 3.29).

In practice, one can use logit or probit as they wish; makes very little difference is terms of substantive results...

Here's an example, using some fake data:

```
. set obs 10
obs was 0, now 10

. gen ystar=invnorm(uniform())

. gen y=0

. replace y=1 if ystar>0
(7 real changes made)

. gen x=ystar+0.5*invnorm(uniform())

. list
          ystar          y          x
  1.   -1.09397          0  -.1020176
  2.   .3670809          1  -.0591711
  3.    .145398          1   -.141263
  4.   .2657781          1  -.1117685
  5.   .4794085          1   .6215442
  6. -1.233643           0  -1.952093
  7.   .3014338          1   .3728438
  8. -1.545905           0  -1.261479
  9.   .1389086          1  -.1914251
 10.   1.133268          1   .7352839
```

```
. probit y x

Probit estimates                                  Number of obs   =         10
                                                  LR chi2(1)      =       6.94
                                                  Prob > chi2     =     0.0084
Log likelihood = -2.6367262                       Pseudo R2       =     0.5684

------------------------------------------------------------------------------
        y |   Coef.   Std. Err.     z   P>|z|    [95% Conf. Interval]
-------------+----------------------------------------------------------------
        x | 2.860457  3.114063     0.92 0.358   -3.242994    8.963909
    _cons | 1.191137   .7305608    1.63 0.103    -.240736     2.62301
------------------------------------------------------------------------------

. logit y x

Logit estimates                                   Number of obs   =         10
                                                  LR chi2(1)      =       6.90
                                                  Prob > chi2     =     0.0086
Log likelihood = -2.6599858                       Pseudo R2       =     0.5646

------------------------------------------------------------------------------
       y |    Coef.   Std. Err.    z   P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       x |  5.239133 6.341525     0.83 0.409    -7.190027    17.66829
   _cons |  2.055426 1.356075     1.52 0.130     -.602432    4.713283
------------------------------------------------------------------------------
```

Note several things...

- The t-scores, levels of significance, model fits, pseudo–R^2s, and even the log–likelihoods are all almost exactly the same, but...
- The logit estimate for $\hat{\beta}_x$ is $\frac{5.239}{2.860} = 1.83$ times as large as the probit one (and the constant terms are also different to a similar degree).

Moreover,

```
. logit ...

. predict loghat
```

```
(option p assumed; Pr(y))

. probit ...

.predict probhat
(option p assumed; Pr(y))

. corr loghat probhat
(obs=10)

             |  probhat   loghat
-------------+------------------
     probhat |   1.0000
      loghat |   1.0000   1.0000
```

...the predictions for the two models are *identical*.

That said, there are still some good reasons for preferring the logit model over the probit model (in my opinion, anyway). A few of these have to do with the fact that there are some easy, intuitive techniques for making logit results comprehensible to a reader that aren't available for the probit model; we'll talk more about that next week...

Additional References

Nagler, Jonathan. 1994. "Scobit: An Alternative Estimator to Logit and Probit." *American Journal of Political Science* 38(1):230-55.

Royston, Patrick, Douglas G. Altman, and Willi Sauerbrei. 2006. "Dichotomizing Continuous Predictors in Multiple Regression: A Bad Idea." *Statistics in Medicine* 25:127-41.