MLE: Categorical and Limited Dependent Variables

Unit 1: Models for Dichotomous Dependent Variables
1. Logit, Probit, and the Generalized Linear Model

PS2730-2021

Week 1

Professor Steven Finkel

Modeling a Binary (Dichotomous) Dependent Variable

- With a dichotomous dependent variable (coded as 0/1),we model the probability that Y=1, or P(Y=1). This follows from the idea of regression as modeling the Expected Value or the conditional mean of Y, given the Xs:

$$E(Y|X) = X\mathrm{B}, \quad where$$

$$X\mathrm{B} = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \ldots \beta_k \mathrm{X}_k$$

- This means that the average value of Y, given the Xs, is a linear function of the Xs.
- What is the "Average" Value of a Dichotomous Variable?
- If you have J cases on 1 and N-J cases on 0, then:

$$E(Y) = (J*1 + (N-J)*0)/N$$

$$E(Y) = J/N$$

- Which is equal to the proportion of cases on 1, or the probability that Y=1, or P(Y=1)

- More formally, dichotomous dependent variable models are based on the *Bernoulli distribution*, which is the discrete probability distribution for a binomial variable in a *single* trial. (In multiple trials the "binomial distribution" represents the number of "successes" (1) in a sequence of independent trials.) The single-trial outcome for Y is distributed as:

$$Y = \begin{Bmatrix} 1, & \pi \\ 0, & 1\text{-}\pi \end{Bmatrix}$$

Where **π** is P (Y=1). So in a Bernoulli distribution we observe the 1s with a P(Y=1) of **π** and the 0s with a P(Y=0) of 1- **π** .

The goal for dichotomous DV models, then, is to model **π** or P(Y=1) as some function of the independent variables

- An initial formulation would borrow directly from the linear regression model:

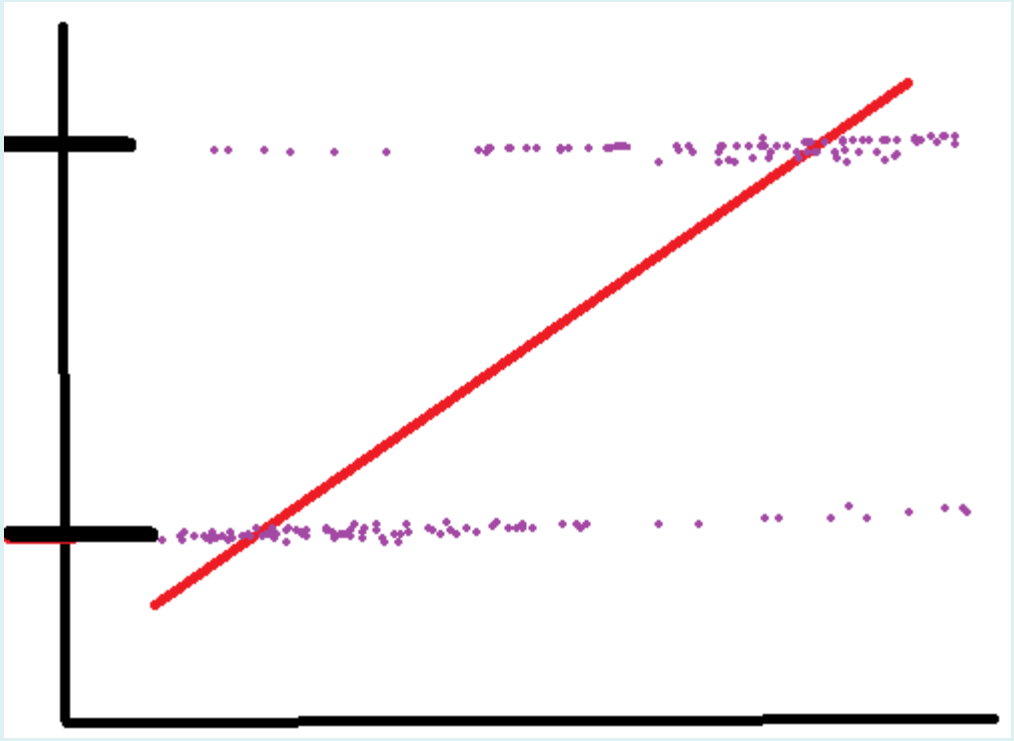
  $$P(Y=1|X)=\pi=X\mathrm{B}$$

  in the bivariate case:

  $$P(Y=1|X_1)=\pi=\beta_0+\beta_1 X_1$$

- This is called the **"Linear Probability Model," (LPM)** as we assume that P(Y=1) is a linear function of the Xs – the effect of a unit change in X will have the same effect on P(Y=1) regardless of where on X the change takes place.
- This is just a regular old regression with the 0/1 Dichotomous variable as the dependent variable, predicted from a series of X independent variables
- Interpretation: For every unit change in X, the probability of Y being equal to 1 increases on average by β units

The Linear Probability Model (LPM)



Estimating the LPM

- Can we estimate the LPM with OLS?
- Error Term is odd: at any value of X_i, there are only 2 values for ε
  - If Y=0, then $0=XB+\varepsilon$, and $\varepsilon = -XB$
  - If Y=1, the $1=XB+\varepsilon$, and $\varepsilon = 1-XB$
  - So ε is *not* normally distributed
  - But with large N that still will not adversely affect estimation
- OLS would still be unbiased, no reason to think that, on average, $E(\varepsilon)\neq 0$, nor that $E(X\varepsilon)\neq 0$ (i.e., no reason to suspect endogeneity simply on the basis of the model choice itself)
- However, OLS will be *inefficient*, as there is intrinsic heteroskedasticity in the model
- You can see this from the graph: there is greater error variance at middle values of P(Y=1) than when P(Y=1) is very large or very small

- This corresponds to the variance of a dichotomous (Bernoulli) variable, which is P(1-P), or π (1- π).
- So: Var(ε)=Var(Y|X)=P(Y=1|X)(1-P(Y=1|X)=XB(1-XB)
- This quantity is largest when P(Y=1|X)=.5, as can be seen from the graph, and will be small when, e.g. P(Y=1|X)=.1 or .9
- This problem can be overcome with a Weighted Least Squares procedure attributed to Arthur Goldberger in the 1960s
- Steps in Goldberger WLS
  - Estimate the LPM with OLS, generate the predicted probabilities $\hat{Y}_i$
  - Calculate weights as $\hat{w}_i = \sqrt{\frac{1}{X\hat{\mathrm{B}}(1-X\hat{\mathrm{B}})}}$ or $\sqrt{\frac{1}{\hat{Y}(1-\hat{Y})}}$
  - Run a weighted regression as:

$$\hat{w}_i Y = X\mathrm{B}\hat{w}_i + \hat{w}_i \varepsilon$$

  - Variance of this new error term is 1, so homoskedastic

Problems with LPM

- Possible predicted P(Y=1) outside of the 0-1 range of logical probabilities. There is no constraint or bound on Y in the LPM
- This affects the first stage of Goldberger's WLS procedure also, and would invalidate the construction of *w* for any case with Y-hat greater than 1 or less than 0, since there would a negative value in the denominator of *w* (no square root possible)
- Most important: theoretically it may not be the case that X has a constant effect on P(Y=1), rather there may be marginal decreasing effects on X as the prior P is very high or very low (e.g., one additional year of graduate school has less effect on voter turnout than changing from no high school degree to one year of college).

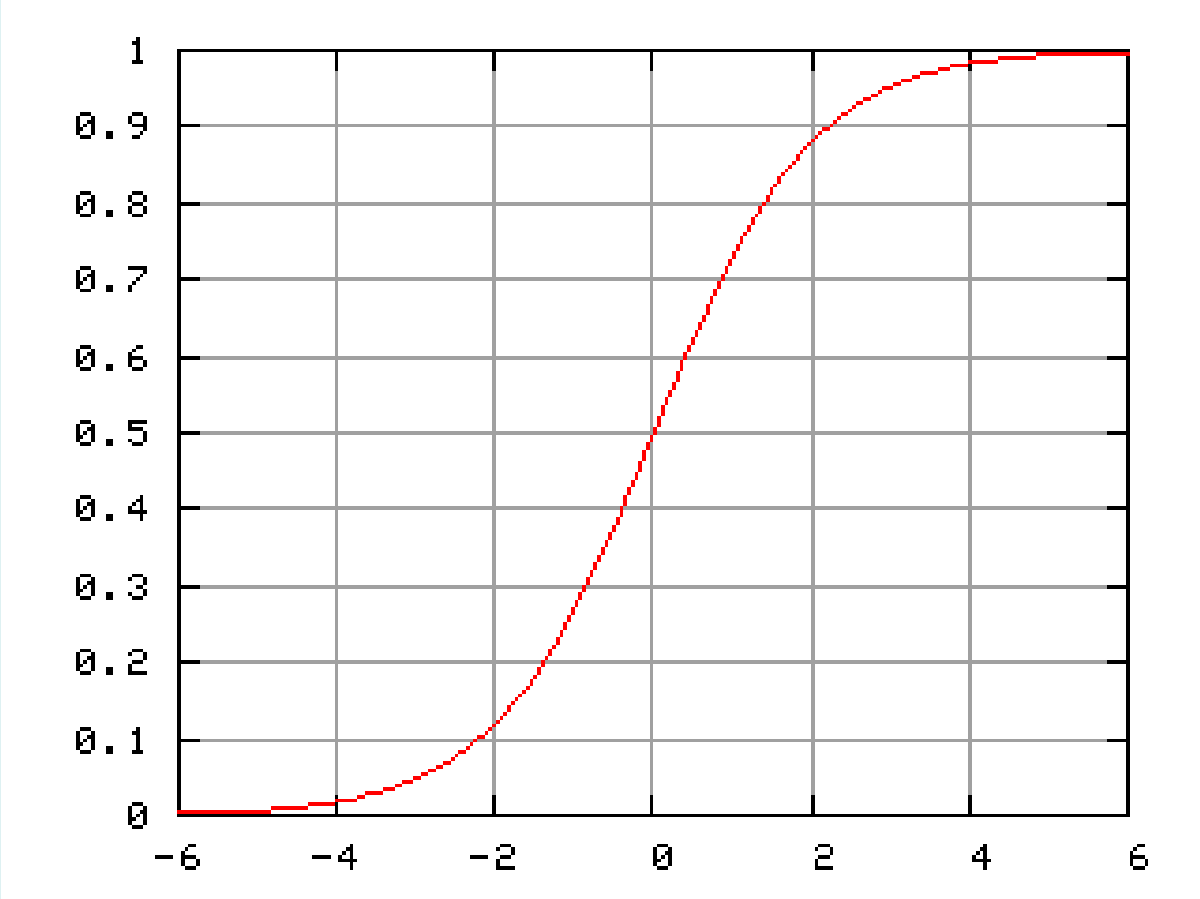
- This is an issue with the **functional form** of the LPM

So we want a model that has a **non-linear functional form** of the effects of X on the P(Y=1) or **π**, where:

- the P(Y=1) are bounded by 0 and 1
- the Xs are unbounded, i.e., can take on any value
- the effects of the Xs are greater at middle levels of the distribution than at the tails

This is the justification for using the LOGISTIC FUNCTION -- or some other "sigmoid" function ("S-shaped") such as the cumulative standard normal distribution, as in probit analysis -- as the basic functional form of a binary or dichotomous variable model

The (Cumulative) Logistic Function



The Logit Model

$$P(Y=1 \mid X) = \frac{\exp(X\mathrm{B})}{1+\exp(X\mathrm{B})} = \frac{1}{1+\exp^{-X\mathrm{B}}}$$

So as XB goes to ∞, P(Y=1) goes to 1 but never gets there;

as XB goes to -∞, P(Y=1) goes to 0 but never gets there;

when XB is 0, P(Y=1)=.5.

So we have a perfectly symmetrical but non-linear functional form with the nice theoretical properties we wanted

Estimation of the Logit Model

- It can be shown that the probability of Y being "0" or (1-P(Y=1)) $= \dfrac{1}{1+\exp(X\mathrm{B})}$
- Given this, we can construct the quantity P(Y=1)/P(Y=0) -- what is called the **"odds"** of Y being 1 -- as:

$$\frac{P(Y=1)}{P(Y=0)} = \frac{\dfrac{\exp(X\mathrm{B})}{1+\exp(X\mathrm{B})}}{\dfrac{1}{1+\exp(X\mathrm{B})}} = \exp(X\mathrm{B})$$

- And taking the natural logarithm of both sides (to the base "e") gives:

$$Ln\frac{P(Y=1|X)}{P(Y=0|X)} = (X\mathrm{B}) = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \ldots \beta_k \mathrm{X}_k$$

We call the log of the odds that P(Y=1) the **"logit"** of Y, and so we can say that the logit model is *linear in the logits*, such that an increase of a unit in X produces a constant change in the *logits*, but is non-linear in the probabilities (and odds). This is how we interpret the estimated β effects, as **linear effects of a unit change in X on the change in the log-odds that P(Y=1)**.

(But P(Y=1) and X are *not* linearly related – IMPORTANT!)

Example of Logits and Probabilities

Summary for variables: locdich
by categories of: groups

groups	mean
0	.2857143
1	.3700441
2	.5311005
3	.7256098
4	.7628866
5	.8253012
Total	.5819149

.

- DV: Dichotomized Local-level Political Participation (YES/NO)
- If no group memberships
  - Probability = .286
  - Odds = .286/.714=.40
  - Log-odds (logit)= ln(.40) = -.92
- If 5 groups
  - Probability=.825
  - Odds=.825/.175=4.714
  - Log-odds (logit)=ln(3.35)=1.55
- All log-odds (logits) less than 0 mean probabilities less than .5 (and odds <1)
- All log-odds (logits) more than 0 mean probabilities greater than .5 (and odds>1)
- Logistic regression models the logits as a linear function of the Xs, using maximum likelihood estimation methods

Bivariate Logistic Regression

```
. logit locdich groups

Iteration 0:   log likelihood = -638.88641
Iteration 1:   log likelihood = -569.38666
Iteration 2:   log likelihood = -568.93407
Iteration 3:   log likelihood = -568.93399
Iteration 4:   log likelihood = -568.93399

Logistic regression                             Number of obs     =        940
                                                LR chi2(1)        =     139.90
                                                Prob > chi2       =     0.0000
Log likelihood = -568.93399                     Pseudo R2         =     0.1095

------------------------------------------------------------------------------
     locdich |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      groups |   .5438632   .0501937    10.84   0.000     .4454854    .642241
       _cons |  -.9634581   .1334391    -7.22   0.000    -1.224994   -.7019222
------------------------------------------------------------------------------
```

- The logit coefficient here for the effect of "group memberships" on "engaged in local participation" is .54.
- This means: for every additional group membership, the log-odds, or logit, of local participation changes, on average, by .54.
- This is a constant linear effect – it is the same when changing from 0-1 exposures, 1-2, 2-3, 3-4, etc.

- To convert into probabilities: we take the predicted logit for an individual with a given value of X, and plug it into the P(Y=1) expression for the logit model:

$$P(Y=1|X)=\frac{\exp(X\mathrm{B})}{1+\exp(X\mathrm{B})}$$

- So when X=0, the predicted logit: -.96 exp(-.96)/(1+exp(-.96))= .28

  X=1, the predicted logit:-.42 exp(-.42)/(1+exp(-.42))=.40

  X=4, the predicted logit: 1.21 exp(1.21)/(1+exp(1.21))=.77

  X=5, the predicted logit: 1.75 exp(1.75)/(1+exp(1.75))=.85

- A unit change in X from 0-1 leads to a .12 change in predicted P(Y=1)

  A unit change in X from 4-5 leads to a .08 change in predicted P(Y=1)

Same unit change in the logit, different unit change in the P(Y=1)!!

The Latent Variable Approach to Modeling Binary Dependent Variables

- Derivation of the logit model was done so far from the need for a non-linear probability model that was bounded by 0,1 with no bounds on the Xs
- Another way of deriving the non-linear functional form for predicting 0,1 dependent variable is based on the "latent variable approach". This usually ends up with the "probit" specification which makes use of the normal distribution (though one could also specify a logistic distribution using this approach and arrive again at the logit model)

- Idea is that you have a 0,1 **observed** variable: vote or not vote; protest or don't protest, war/no war but there is an underlying, latent **"propensity"** to vote, to protest, to go to war which is a continuous, unobserved variable.
- So can imagine that the latent "propensity" variable might run from negative infinity to infinity, and that there is some threshold point beyond which we observe a voter, a protester, or a conflict. So can view the observed 0,1 variable as **mapped from a continuous latent, unobserved variable that has no bounds.**
- This also fits the notion of **"Expected Utility"** models of behavior perfectly: the utility derived from one behavioral choice versus another can be infinitely negative or positive, and at the threshold of (for example) zero you observe behavior "1", and below the threshold you observe behavior "0"
- Many discrete choice models we'll examine relating to ordinal and multinomial outcomes are derived from this framework

- Model:

$$Y_i^* = \Sigma\beta X_i + \varepsilon_i$$

$$Y_i^* = X\mathrm{B} + \varepsilon_i \quad \text{in matrix type notation}$$

$$\mathrm{E}(Y_i^* \mid X) = X\mathrm{B}$$

- Y* is a continuous *unobserved* variable. It is mapped to the observed dichotomous variable Y through a "measurement equation" that says if Y* is above a certain threshold τ, then the observed Y will be 1; if Y* is below the threshold, then observed Y will be 0

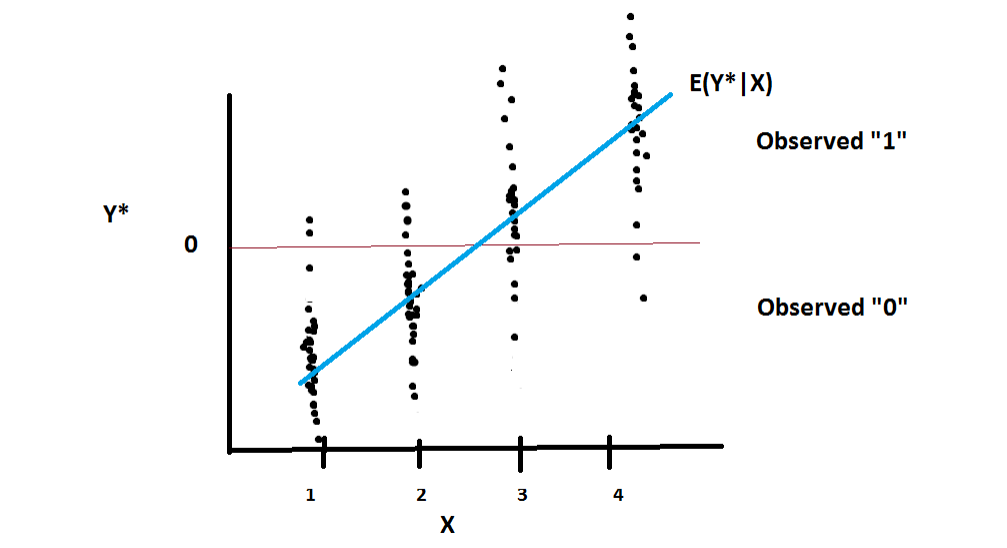
$$Y_i = 1 \text{ if } \mathrm{Y}_i^* > \tau$$

$$Y_i = 0 \text{ if } \mathrm{Y}_i^* \leq \tau$$

- Assuming τ to be 0 (following the logic above), then

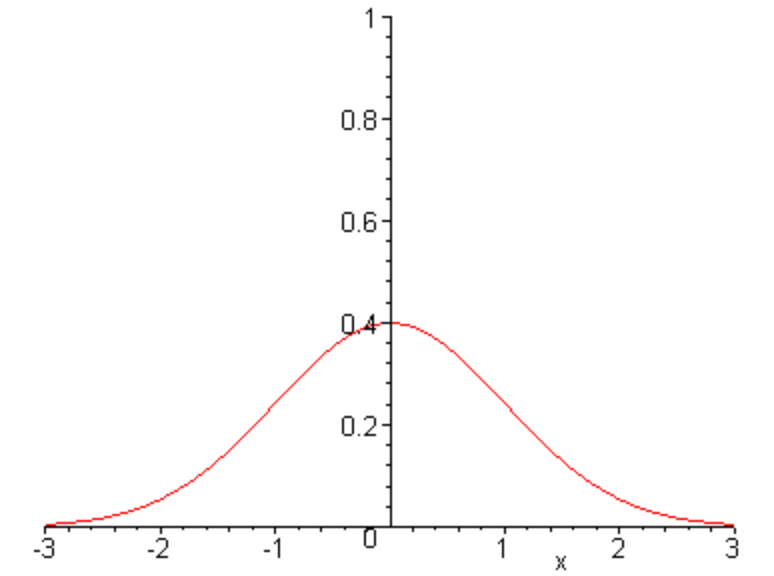
$$Y_i = 1 \text{ if } \mathrm{Y}_i^* > 0$$

$$Y_i = 0 \text{ if } \mathrm{Y}_i^* \leq 0$$



- Y* is unobserved, so we can't estimate with OLS. We use ML methods, which we will introduce next time. For now, need to make some assumptions about the error term ε in this model in order to identify the model parameters
- Assume that ε is a standard normal variable with mean of 0 and variance of 1
- This identifies the variance of Y*, which is unobserved. (And we can arbitrarily make this assumption with no substantive implications – it only changes the relative value of the regression coefficient but not the substantive relationship, though there are some potentially problematic implications we'll discuss)
- So $\varepsilon \sim N(0,1)$
- We could also say that ε is distributed logistically. Then $var(\varepsilon)=\pi^2/3$

The Standard Normal Distribution

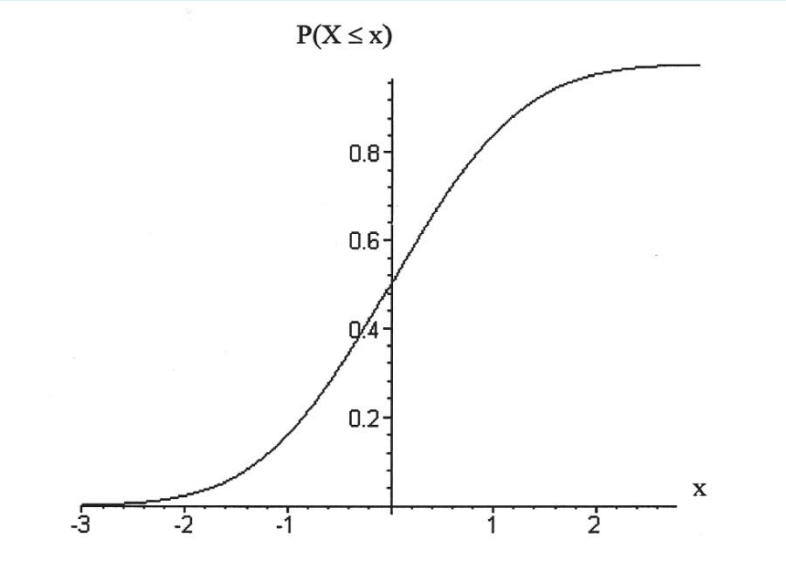


$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-u}{\sigma}\right)^2}$$

for $-\infty < x < \infty$

- This is a graph of the probability density function or pdf of the values in a standard normal variable.
- On the X axis is the "z-score" of *any* normally distributed variable
- The f(x) gives you the height, or "pdf" of the curve at any given point or X value (z-score) on the variable
- There is actually not a probability value associated with observing X at any single discrete point since this is a continuous function, but there are probabilities associated with observing Xs *between* two points on the pdf

The Cumulative Normal Distribution Function



z-score

Probit models will give coefficients that indicate the average change in the z-score of Y for a given unit change in X, and this translates into changes in the probability that Y=1 via the cdf

- The cumulative normal distribution function (the "cdf") gives the proportion of the standard normal curve at or below a given value
- So a z-score of 0 corresponds to a cdf of .5; this means that .5 of the normal curve is at or below a z-score of 0
- A z-score of -1 corresponds to a cdf value of .16, or 16% of the function
- A z-score of 1 corresponds to a cdf value of .84. or 84% of the function
- These cdf values correspond also to the **cumulative probabilities** of observing values in the distribution at or below the given value. So a z-score of 1 has a cumulative probability of .84, e.g.
- We represent this as, e.g., $\Phi(0)=.5$
- $\Phi(-1)=.16$ $\Phi(1)=.84$ 16 $\Phi(.5)=.69$

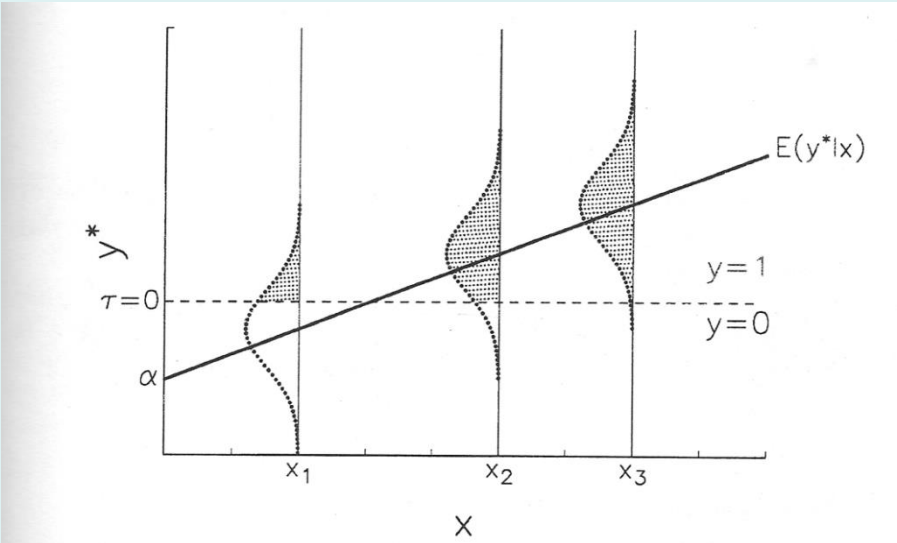


Figure 3.2. The Distribution of y^* Given x in the Binary Response Model

- Back to our probit derivation. Assuming normally distributed errors in the equation for Y*:
- Can see that sometimes, for a given level of X (and thus a given value of predicted Y*), the person will have an observed value of 1 if the error term is sufficiently large to push her over the threshold τ (0). If not, we observe 0.
- Can also see that the probability of obtaining an error term large enough to push a person over the threshold is greater, as X increases (given the positive relationship here between XB and Y*). When X=3, e.g., it will be very unlikely to get an error sufficiently negative to push the person *under* the τ (0) threshold, so the probability of observing a 1 will be very high
- But even if Y* is very low (high), there is still a chance that a very high (low) ε leads to an observed 1 or 0.
- With the assumption of a normally distributed error term, can calculate the probabilities exactly!

The Probit Model

$$Y^* = X\mathrm{B} + \varepsilon$$

$$P(Y = 1 \mid X) = P(Y^* > 0 \mid X)$$

$$P(Y = 1 \mid X) = P(X\mathrm{B} + \varepsilon > 0 \mid X)$$

$$P(Y = 1 \mid X) = P(\varepsilon > -X\mathrm{B} \mid X)$$

- So the probability that Y=1 is equal to the probability of obtaining an error term greater than –XB, which will push (or keep) the person over the threshold of 0
- If XB puts the person at -1, for example, we will observe a "1" only if the error term is greater than 1, which would make Y* greater than 0.
- Given normal curve probabilities, we know that will happen with .16 probability. (How?)
- If XB puts the person at 1, we will observe a "1" if the error term is greater than -1. We know that will happen with .84 probability

- We want to know the probability that $\varepsilon > -XB \mid X$
- Notice, though, that the probability of observing an error term *greater than* $-XB$ is the same as the probability of observing an error term *less than* XB; this follows from the symmetry of the normal distribution
  - For an XB of 1:
    - Probability that $\varepsilon > -1$ (from previous slide) is .84
    - Probability that $\varepsilon > 1$ (from previous slide) is .16, so Probability that $\varepsilon < 1$ is also .84
  - For an XB of -1.5:
    - Probability that $\varepsilon > 1.5$ is .064
    - Probability that $\varepsilon > -1.5$ is .934, so Probability that $\varepsilon < -1.5$ is .064
- So Probability that $\varepsilon > -XB$= Probability that $\varepsilon < XB = \Phi(XB)$
- So the Probability of obtaining an error term greater than -XB – which will push Y* above the threshold of 0 so that observed Y will be 1, is equal to the cumulative probability (or proportion of the cumulative normal distribution) associated with XB
- This is the probit model! $P(Y=1 \mid XB) = P(\varepsilon > -XB) = \Phi(XB)$

- The probit model for binary dependent variables:

$$P(Y=1 \mid X) = \Phi(X\mathrm{B})$$

- The probability that Y=1 is equal to 1 is equal to the cdf – the value of the cumulative normal distribution function – associated with the z-score quantity XB
- Probit is "linear in the z-scores" and non-linear in the probabilities, just like logit was "linear in the logits" and non-linear in the probabilities

Examples of Probit XB and P(Y=1). In STATA: display normal(XB)

- Model: $Y^*=-1+.2*X$
  - X=-1 XB=-1.2 $P(Y=1)=\Phi(-1.2)=.12$
  - X=0 XB=-1 $P(Y=1)=\Phi(-1)=.16$
  - X=5 XB=0 $P(Y=1)=\Phi(0)=.5$
  - X=10 XB=1 $P(Y=1)=\Phi(1)=.84$
- Change intercept to +1: $Y^*=1+.2X$
  - X=-1 XB=-.8 $P(Y=1)=\Phi(.8)=.79$
  - X=0 XB=1 $P(Y=1)=\Phi(1)=.84$
  - X=5 XB=2 $P(Y=1)=\Phi(2)=.98$
  - X=10 XB=3 $P(Y=1)=\Phi(3)=.99$
- Change slope to .5: $Y^*=-1+.5X$
  - X=-1 XB=-1.5 $P(Y=1)=\Phi(-1.5)=.07$
  - X=0 XB=-1 $P(Y=1)=\Phi(-1)=.16$
  - X=5 XB=1.5 $P(Y=1)=\Phi(1.5)=.93$
  - X=10 XB=4 $P(Y=1)=\Phi(4)=.99$

Summary of Logit/Probit Models

$$P(Y=1|X)=\Phi(X\mathrm{B}) \qquad \text{Probit}$$

$$P(Y=1|X)=\frac{\exp(X\mathrm{B})}{1+\exp(X\mathrm{B})} \qquad \text{Logit}$$

- These are both non-linear probability models bounded by 0 and 1
- In each case, though, *something* has a linear relationship to X as well
  - Probit: The Y and X relationship is linear in the z-scores
  - Logit is linear in the log-odds
- We can go from probabilities to z-scores or log-odds and back again via the "inverse" of the P(Y=1) functions above

$$X\mathrm{B}=\Phi^{-1}P(Y=1|X) \qquad \text{Probit}$$

$$X\mathrm{B}=\ln\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) \qquad \text{Logit}$$

- This leads to a general approach to many probability models called **"Generalized Linear Models"** and unifies much of what we will do in the class
- Start with the assumed probability distribution for Y
- Some possible distributions we'll consider:

Normal (as in slide 21)

$$f(Y \mid \mu) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{Y-\mu}{\sigma}\right)^2}$$

Binomial (number of successes y in n trials), or Bernoulli one trial

$$f(y \mid n, \pi) = \left(\frac{n}{y}\right) \pi^y (1-\pi)^{n-y} \qquad f(y, \pi) = \pi^y (1-\pi)^{1-y}$$

Poisson (number of occurrences in a specific time interval, given a rate parameter μ):

$$f(y \mid \mu) = \mu^y \frac{e^{-\mu}}{y!}$$

- Then, define a linear prediction as a function of the independent variables X

$$\eta = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_q X_q = XB$$

- Finally, relate the expected (mean) response probability (***μ***) or (***π***) or **P(Y=1|X)** of the original distribution to the linear prediction via what is called a "link function" – i.e., go from the potentially non-linear response probabilities to the linear prediction

$$g(\mu) = \eta = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_q X_q = XB$$

- Where the link "g" will differ depending on the probability distribution at hand, e.g. normal, binomial (Bernoulli), poisson, etc., but the general unifying principle is the same

- Linear (identity) link: $g(\mu) = g(\pi) = \eta = XB$
- Logit link:

$$g(\mu) = g(\pi) = ln\frac{\pi}{(1-\pi)} = \eta = XB$$

- Probit link:

$$g(u) = g(\pi) = \Phi^{-1}(\pi) = \eta = XB$$

- Poisson link:

$$g(\mu) = \ln(\mu) = \eta = XB$$

- Can also go in the reverse direction, from the linear predictor to the mean response probability, via the "mean" function or the *inverse* of the link function

$$\mu = g^{-1}(\eta)$$

- Linear mean function: $\mu = \pi = g^{-1}(\eta) = \eta = XB$

- Logit mean function: $\mu = \pi = g^{-1}(\eta) = \dfrac{e^{\eta}}{1+e^{\eta}} = \dfrac{1}{1+e^{-\eta}}$

- Probit mean function: $\mu = \pi = g^{-1}(\eta) = \Phi(\eta)$

- Poisson mean function: $\mu = g^{-1}(\eta) = e^{\eta} = e^{XB}$

- All these models have similar form: they conceptualize the effect of Xs on E(Y) with a linear component in the Xs, and then some link between that linear component and the potentially non-linear E(Y|X)
- All estimated via ML methods we'll consider next time
- All unified under the GLM routine in most software packages (Stata, R, e.g.)
- Specification: distribution (binomial, e.g.); link function (logit or probit, e.g.)
- Sometimes will estimate two parameters: mean **and** variance of the response distribution
- Can further generalize the GLM to analysis of multilevel and longitudinal data, adding random effects in $\boldsymbol{\eta}$ at higher levels of the data hierarchy to account for clustering, as we will see later in the course