ECON B253: INTRODUCTION TO ECONOMETRICS

Sebastian Anti, Assistant Professor of Economics
Bryn Mawr College

Lecture 6: Binary Dependent Variable Models

Dichotomous, Binary, or "Dummy" Dependent Variables

- **Many outcome variables we want to study are dichotomous, or binary. Examples:**
  - School attendance
  - Labor force participation
  - Experience of hunger
- **As with an indicator on the RHS, data codes these responses as follows:**
  - = 1 if individual attends school in a given year, = 0 if not
  - = 1 if individual participates in the labor force, = 0 if not

Statistics Review

- Let Y be a binary random variable that can take the value of 0 or 1
- The probability density function of Y can be written in the simple table:

Possible Value (v)	Probability $Y = v$
0	$1 - p$
1	p

- Examples are a fair coin toss or choosing a child at random from a population and observe whether or not they attend school

Main Estimation Techniques for Binary Dependent Variables

- **Traditionally 3 main approaches:**
  1. **OLS: "Linear Probability Model" (LPM)**
     - In fashion right now. Younger researchers seem to prefer it for several advantages we'll discuss
  2. **Probit**
     - Traditional estimator. Historically very popular but seems to be loosing popularity recently
  3. **Logistic Models or "Logits"**
     - Also traditionally popular. Loosing popularity in recent years

Linear Probability Models

- **Linear probability models (LPM) is just using OLS with a binary dependent variable**
- **What happens when Y is a dichotomous dependent variable?**

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i, and \tag{9}$$

$$E(u_i | X_i, Z_i) = 0 \; \forall \; i \tag{10}$$

- **It's a little hard to think of a unit change in X leading to a unit change in Y when Y can only take two values!**

Linear Probability Models

- **We can re-express the model as follows,**

$$E(Y_i|X_i, Z_i) = \beta_0 + \beta_1 X_i + \beta_2 Z_i \qquad (11)$$

- **Recalling the properties of binary random variables, this is equivalent to the following,**

$$P(Y = 1|X_i, Z_i) = \beta_0 + \beta_1 X_i + \beta_2 Z_i \qquad (12)$$

- **How can we think about what we are estimating when we run the regression and estimate the betas?**
  - How are probabilities measured?

Linear Probability Models

- **We're thinking that the probability that $Y = 1$ differs across groups, which differ in values of X and Y**
- **β_1 will tell us the change in probability of $Y = 1$ when X increases by 1 unit**
  - If $\beta_1 = 0.07$, then a one-unity change in X increases the probability of $Y = 1$ by 7 ***percentage points***
  - Note how this is different than saying "increases by 7 percent"

Advantages and Disadvantages of LPMs

1. **Advantage: Coefficients are easy to interpret as changes in percentage points of a probability**
   - Much easier than Probits and Logits (more on this in a couple slides)
2. **Advantage: Computationally easier, allowing for high dimensional fixed effects**
   - More on this in a couple weeks
3. **No problematic assumptions on the normality of the error term (issue with the Probit)**
4. **Disadvantage: Can produce predicted values (probabilities) outside 0 to 1**
   - Probits and Logits don't have this problem

Dealing with the LPM Disadvantage

- **So if we are worried that the LPM is predicting values above 1 and below 0, you might think of constructing a model that is a functional transformation $F(\cdot)$ of the linear model such that**

$$P(Y = 1|X_i, Z_i) = F(\beta_0 + \beta_1 X_i + \beta_2 Z_i) \in [0,1] \qquad (13)$$

  - So $F(\cdot)$ is a function that has some characteristic that limits itself to values of 0 and 1

- **This is the idea behind Probit Models and Logit Models**

The Probit Model

- **Instead of assuming that $P(Y = 1)$ is a linear function of X and Z, assume it is a non-linear function that won't allow values greater than 1 or less than 0,**

$$E(Y|X,Z) = Pr(Y = 1|X,Z) = \Phi(\beta_0 + \beta_1 X_i + \beta_2 Z_i) \qquad (14)$$

Where $\Phi(a)$ is the standard normal cumulative distribution function:
$\Phi(a) = Pr(z < a) = z \sim N(0,1)$

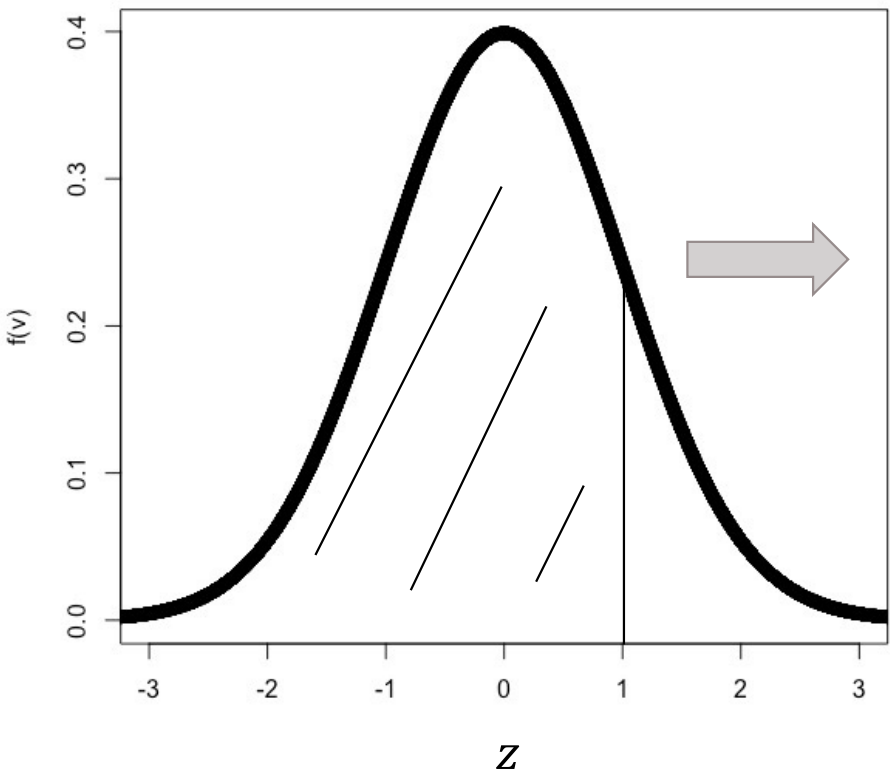
- Takes the form of an integral of the normal PDF up to z: $\Phi(a) = \int_{-\infty}^{z} \phi\,(v)dv$
- What is a cumulative distribution function (CDF)?

The Cumulative Distribution Function

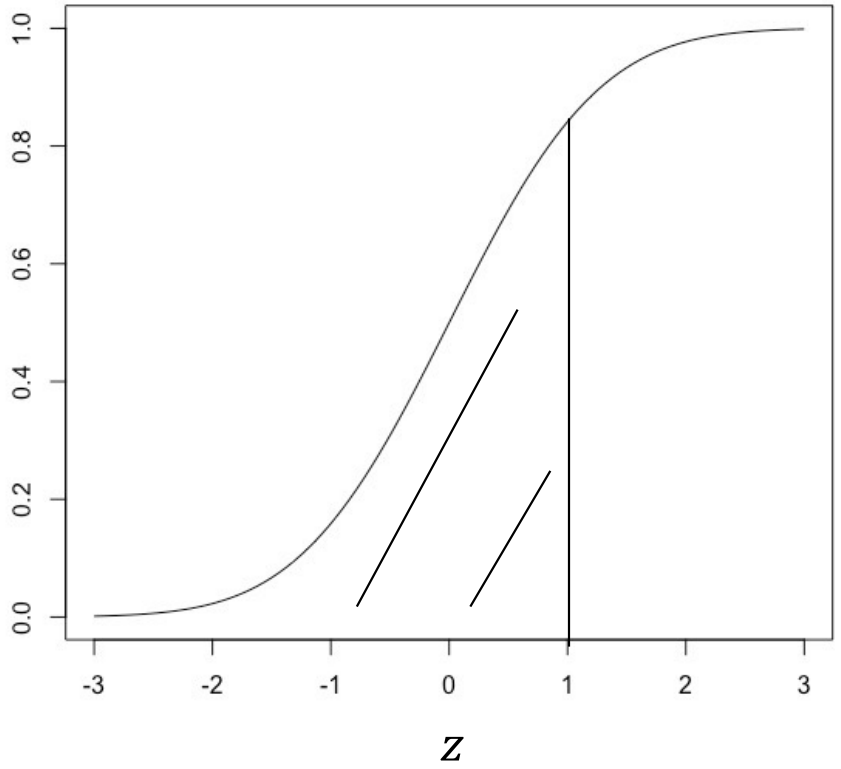
- **Recall the probability density function (PDF) of a random variable Y**
  - It is a statement of the possible values Y can take on along with a statement of the probability Y takes each of those values
- **The cumulative density function (CDF) of a random variable Y is,**
  - The probability of Y being less than or equal to a specified real number

Comparing the Standard Normal PDF to the CDF

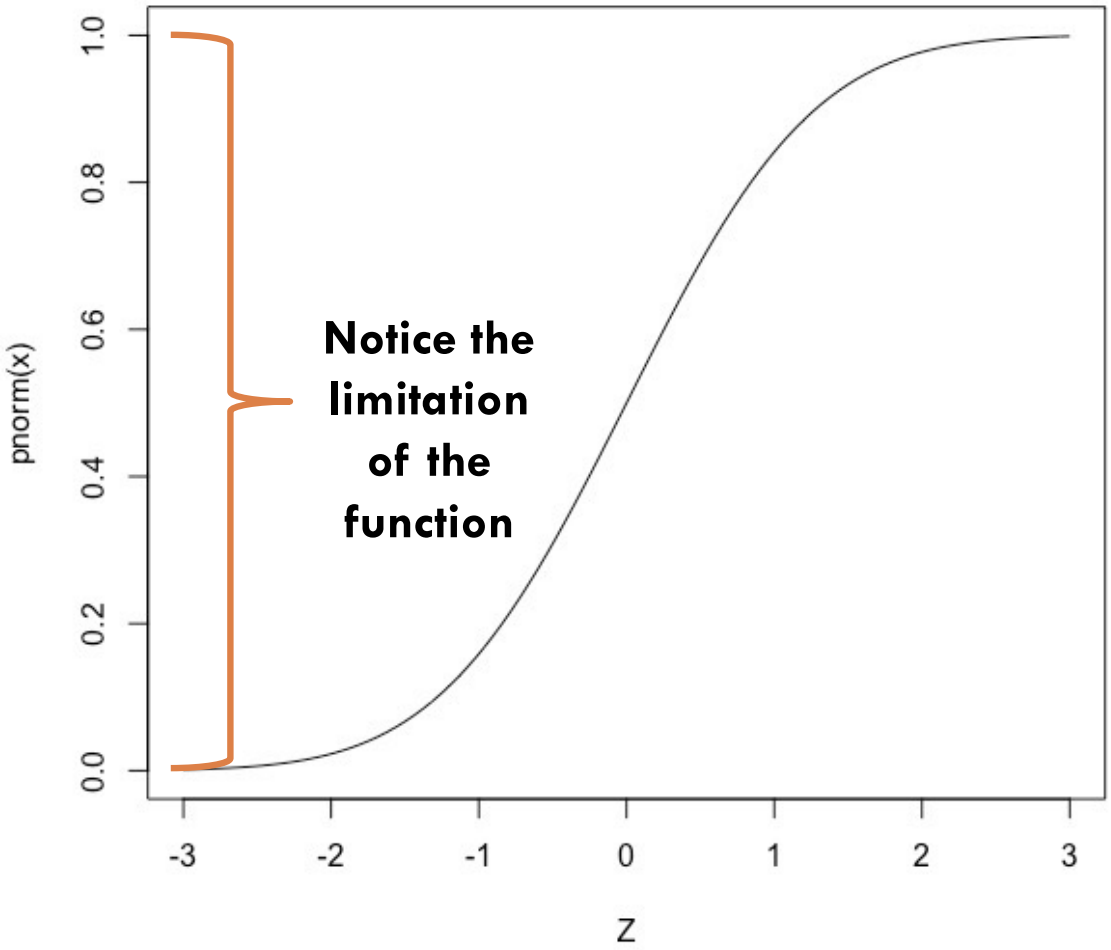
Standard Normal PDF



Standard Normal CDF



The Standard Normal CDF



What is $\Phi(0)$? (0.5)

What is $\Phi(-2)$? (0.05)

What is $\Phi(2)$? (0.95)

The Probit Model

$$E(Y|X,Z) = Pr(Y = 1|X,Z) = \Phi(\beta_0 + \beta_1 X_i + \beta_2 Z_i) \quad (15)$$

Where $\Phi(a)$ is the standard normal cumulative distribution function: $\Phi(a) = Pr(z < a) = z \sim N(0,1)$

- **So note that the Probit model maps to a z score in the standard normal distribution's CDF**
  - This then maps to a probability of $Y = 1$, this is bounded by 0 and 1!
- **Because the dependent variable is the probability of $Y = 1$, the coefficients on the independent variables in a Probit model are in z scores**
  - This makes them hard to interpret on their own and require some extra work to make sense of

How to Estimate the Probit Model?

- **We can't use OLS because the Probit model is non-linear in the parameters**
  - So we need another approach
- **An alternative approach: The Maximum Likelihood (ML) Principle**
  - ML takes the likelihood function, which is the joint probability distribution of the data treated as a function with unknown coefficients. ML then finds the parameter values that would most likely produce that JPDF
    - In other words, ML fits the model to the data in a way that maximizes the probability of getting a dataset that looks like the one we have
- **Why ML?**
  - Many good statistical properties (including asymptotic unbiasedness and efficiency)
  - It can be shown that OLS *is* the maximum likelihood estimator for a standard linear model
  - Requires numerical maximization

Maximum Likelihood for the Probit Model: First derive the Log-Likelihood function

- **Write down the likelihood function:**

$$Pr(Y_1 = y_1, \dots, Y_N = y_N) = Pr(Y_1 = y_1) \times \cdots \times Pr(Y_N = y_N)$$

$$= \left[p_1^{y_1}(1-p_1)^{1-y_1}\right] \times \cdots \times \left[p_N^{y_N}(1-p_N)^{1-y_N}\right] \quad (26)$$

- **Remember that $\boldsymbol{p_i = \Phi(X_i, Z_i) = \Phi(\beta_0 + \beta_1 X_i + \beta_2 Z_i)}$. Substituting this in,**

$$= \left[\Phi(\beta_0 + \beta_1 X_1 + \beta_2 Z_1)^{y_1}\left(1 - \Phi(\beta_0 + \beta_1 X_1 + \beta_2 Z_1)\right)^{1-y_1}\right] \times \cdots$$

$$\times \left[\Phi(\beta_0 + \beta_1 X_N + \beta_2 Z_N)^{y_N}\left(1 - \Phi(\beta_0 + \beta_1 X_N + \beta_2 Z_N)\right)^{1-y_N}\right] \quad (27)$$

- **(Yikes!) Take the log of this expression. By log rules, you can then express this as the summation on the next slide**

Maximum Likelihood for the Probit Model

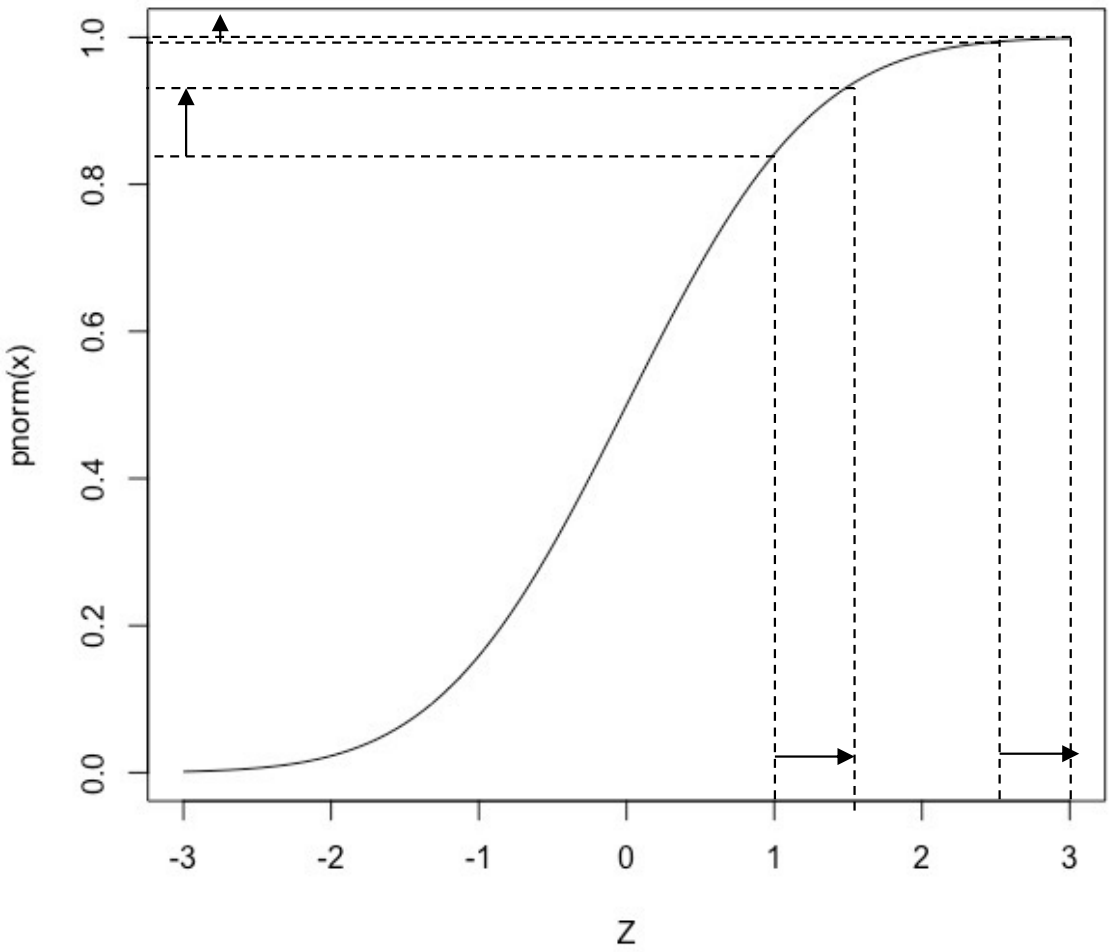
- **Now we need to solve the following maximization problem to find β_0, β_1, and β_2,**

$$Max_{\{\beta_0,\beta_1,\beta_2\}} \sum_{i=1}^{N} Y_i \, ln[\Phi(\beta_0 + \beta_1 X_i + \beta_2 Z_i)]$$

$$+ \sum_{i=1}^{N} (1 - Y_i) \, ln[1 - \Phi(\beta_0 + \beta_1 X_i + \beta_2 Z_i)] \qquad (28)$$

- **No way to do this by hand. You must use a computer**

Coefficient Units are in Z Scores



- The effect of z score change will change based on the values of the independent variables.
- Suppose X and Z are such that z = 1. A .5 increase z scores is as follows,

$$E(Y|X,Z) = Pr(Y = 1|X,Z)$$
$$= \Phi(\beta_0 + \beta_1 X_i + \beta_2 Z_i)$$

Where $\Phi(a)$ is the standard normal cumulative distribution function:

$$\Phi(a) = Pr(z < a) = z \sim N(0,1)$$

Interpreting this Coefficient

- **So what we'd want to do is plug in some numbers for X and calculate how the probability of $Y = 1$ changes**
  - Usually you'll just do that at the means of the independent variables
- **You can do this easily with the command `dprobit` in Stata**
  - These coefficients are interpreted as probabilities

Better Stata Code

- You can do this with the code **`dprobit varY varX`**

```
Probit regression, reporting marginal effects          Number of obs = 108991
                                                       Wald chi2(7)  = 174.88
                                                       Prob > chi2   = 0.0000
Log pseudolikelihood = -51499.724                      Pseudo R2     = 0.0068

                          (Std. Err. adjusted for 4,688 clusters in basin_id)
------------------------------------------------------------------------------
             |               Robust
epide~ea     |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
-------------+----------------------------------------------------------------
ntwk_u..     |   .0380562   .0503572     0.76   0.450   .051898  -.060642  .136754
ntwk_u..     |   .2979763   .1164124     2.56   0.010   .007664   .069812   .52614
ntwk_d..     |   .6411111   .0924214     6.98   0.000   .080519   .459969  .822254
ntwk_d..     |  -.2627018   .1259005    -2.09   0.037   .013534  -.509462 -.015941
popula~n     |   2.25e-06   6.67e-07     3.37   0.001   2231.14   9.5e-07  3.6e-06
ntwk_u..     |  -6.40e-07   3.04e-07    -2.11   0.035    8016.8  -1.2e-06 -4.4e-08
ntwk_d..     |  -7.25e-06   7.62e-07    -9.60   0.000   10381.2  -8.7e-06 -5.8e-06
-------------+----------------------------------------------------------------
  obs. P     |   .1828867
 pred. P     |    .180897  (at x-bar)
------------------------------------------------------------------------------
    z and P>|z| correspond to the test of the underlying coefficient being 0
```

Interval Estimation with Probit Models

- **It is possible to derive formulas for the standard errors of the coefficient estimates and the estimated probability impacts and estimate standard errors**
  - These formulas are very complicated and cumbersome and not worth our time here to go into great detail
  - Stata reports these automatically
    - Doesn't use Student's t distribution, so the key test statistic isn't the t but z. Same rough critical value threshold for statistical significance though
    - This is one of the few times I'll say "just trust Stata"
    - If you need to reference any of these formulas, they're available in any advanced econometrics textbook
- **We use the estimated standard errors and statistical tests in the same way we did with OLS**

Joint Hypothesis Tests in Probit Models

- **We can't use F tests**
  - Why? They're based on the sum of squared residuals which is critical in OLS but has no meaning here
- **Other tests of joint hypotheses are possible**
  - These are not worth out time to outline in detail since they are used in special situations
  - If you encounter them at some point you can find them easily in textbooks. Here are the common ones:
    - Wald test, Likelihood ratio test, and Lagrange Multiplier Test
    - Of these, usually the easiest is the Likelihood Ratio Test
      - Check it out in Wooldridge if your interested

Notational Points with Probits (and logits)

- **Often, researchers will express their model linearly and simply say they are estimating it with a probit (or a logit)**
- **Other times, they'll plug in their model to the standard normal PDF (or logistic function) as follows in some form of the following notation,**

$$P(Y_i) = \pi(Y_i) = \Phi(\beta_0 + \beta_1 X_i + \beta_2 Z_i) \qquad (31)$$

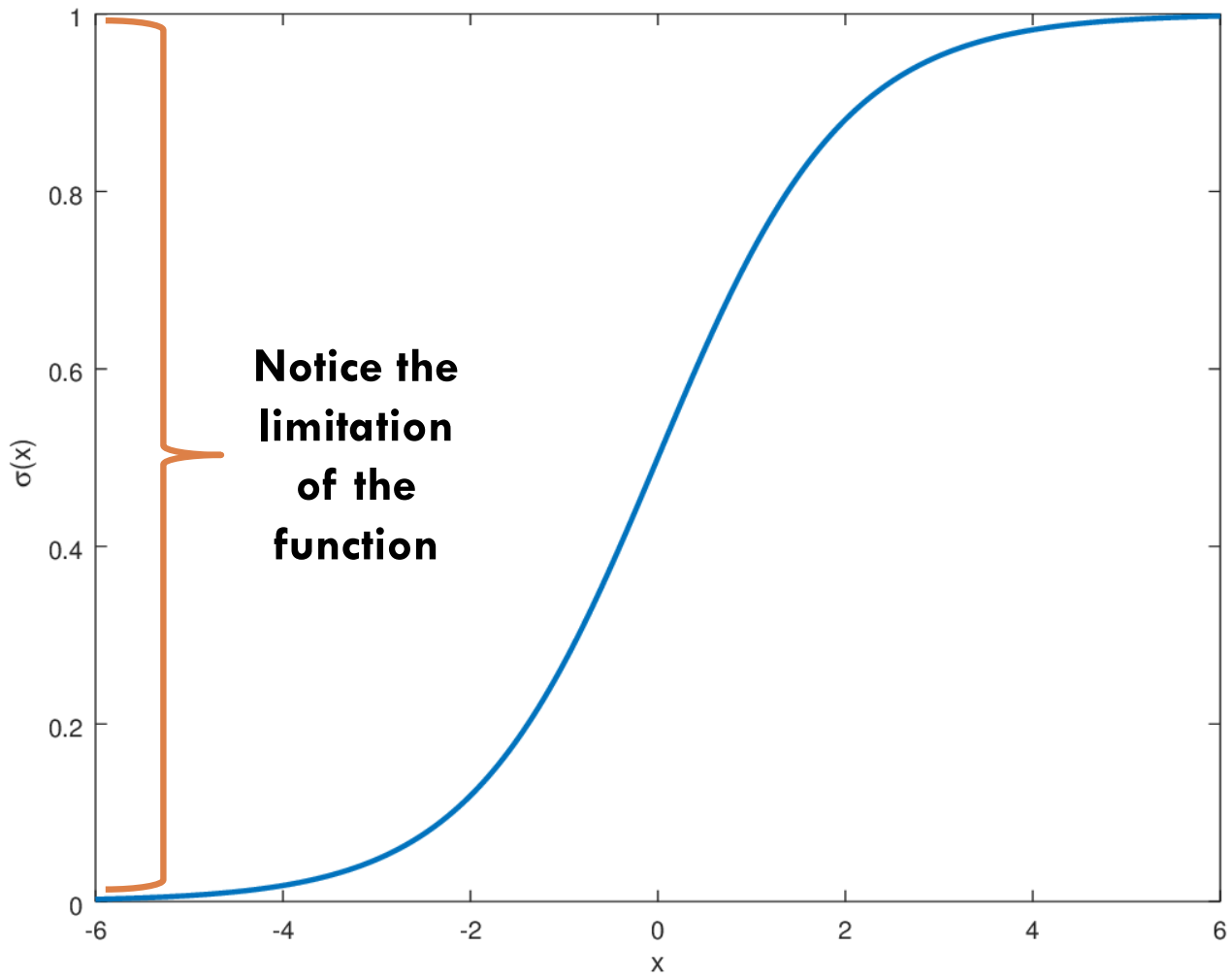
Logit Models

- **A popular alternative to Probit models and LPMs**
  - Also good when the dependent variable is multinomial, either ordered or unordered
- **Instead of assuming that G is a standard normal CDF, we assume that we have a logistic function $L(\cdot)$:**

$$E(Y|X,Z) = Pr(Y = 1|X,Z) = G(\beta_0 + \beta_1 X_i + \beta_2 Z_i)$$
$$= L(\beta_0 + \beta_1 X_i + \beta_2 Z_i) \quad (31)$$

- **Where $L(z) = \frac{e^z}{1-e^z}$, a CDF for a standard logistic random variable**
  - $L(z)$ tends to 0 as z gets very negative and tends toward 1 as z gets very positive

The Logistic Function



Estimating The Logit Model

- Again, good old OLS isn't possible because the coefficients are not linear in the parameters, so…
- The logit model again uses maximum likelihood to estimate the model's parameters

Maximum Likelihood for the Logit Model

- **Now we need to solve the following maximization problem to find β_0, β_1, and β_2,**

$$Max_{\{\beta_0,\beta_1,\beta_2\}} \sum_{i=1}^{N} Y_i \, ln\left[\frac{e^{\beta_0+\beta_1 X_i+\beta_2 Z_i}}{1+e^{\beta_0+\beta_1 X_i+\beta_2 Z_i}}\right]$$

$$+\sum_{i=1}^{N}(1-Y_i)\, ln\left[\frac{1}{1+e^{\beta_0+\beta_1 X_i+\beta_2 Z_i}}\right] \qquad (33)$$

- **Again, no way to do this by hand. You must use a computer**

Interpreting Logistic Regressions

- **Coefficient you get from logit estimates again aren't useful on their own due to interpretation issues**
- **However, it's easy to convert coefficients into "odd's ratios" which is a straightforward measure. You should always do this**
- **Doing this is very straightforward:**
  - All you do is calculate $e^{\widehat{\beta}_1}$, remember $e \approx 2.718$
  - If $e^{\widehat{\beta}_1} > 1$, then a one unit increase in X corresponds with a $e^{\widehat{\beta}_1} - 1$ percent increase in the likelihood of Y
  - If $e^{\widehat{\beta}_1} < 1$, then a one unit increase in X corresponds with a $1 - e^{\widehat{\beta}_1}$ percent decrease in the likelihood of Y

Notational Points with Logits

- **Again, often researchers will express their model linearly and simply say they are estimating it with a logit**
- **Other times, they'll plug in their model to the logistic function as follows in some form of the following notation,**

$$P(Y_i) = \pi(Y_i) = \frac{e^{\beta_0+\beta_1 X_i+\beta_2 Z_i}}{1 - e^{\beta_0+\beta_1 X_i+\beta_2 Z_i}}$$

$$= \frac{1}{1 + e^{-(\beta_0+\beta_1 X_i+\beta_2 Z_i)}} \quad (34)$$

Advantages and Disadvantages of Probits and Logits

1. **Advantage: Ensures predicted values are between 0 and 1**
2. **Disadvantage: Coefficients are hard to interpret**
3. **Disadvantage: Computational issue with high-dimensional fixed effects**
4. **Disadvantage (Probit): Normality of the error term is not a trivial assumption**
   - Advanced point: We didn't study this, but we can get around this assumption with OLS by invoking its asymptotic properties as N gets large. You can't invoke this with the Probit

Current State of the Profession

- **It used to be that if the outcome variable was binary, Probits and Logits were the automatic choice**
  - This has changed as researchers have become much more worried about the assumption of normality of the error term and the use of high-dimensional fixed effects have become common ways to minimize bias
    - Look at Teevrat Garg's work. Heavy use of FE.
  - Also, it is unclear why predicted probabilities outside of 0 and 1 are actually a bad thing. <u>There is no formal statistical reason this is problematic</u>

The Current State of the Profession

- **That said, they are very common and you will see them a lot in the literature**
  - When you read this type of work, you should be aware of these criticisms of this approach
- **ML estimators are particularly still common in trade and investment econometrics. If this interests you and you're thinking about a thesis in these subjects, check out:**
  - The Poisson Estimator (when the dep. var. takes on integer values)
  - The Poisson Pseudo-Maximum Likelihood (PPML) Estimator (standard estimator with "gravity models")
  - The Heckman Selection Estimator (approach to deal with "censored" data)