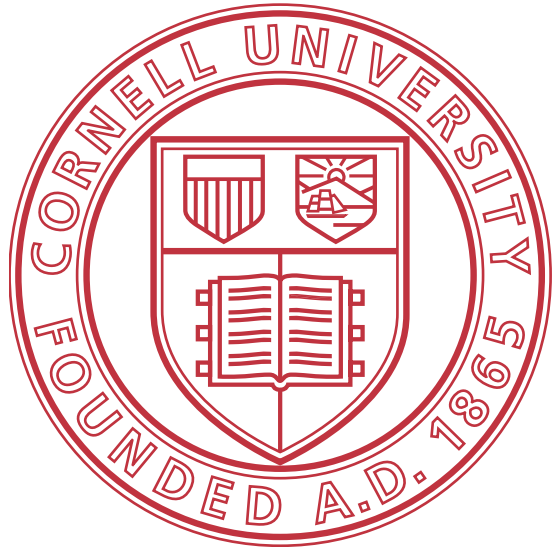
ECON 3140: Introduction to Econometrics
Part 6: Discrete Regressors and Outcomes



March 6, 2025

Regression with Binary Outcomes

We now switch gears and put an indicator variable on the left-hand side.

This is extremely important in economics because many decisions are binary. Economists try to explain:

- entry/exit to and from the labor force,
- entry/exit into and from a specific market,
- fertility decisions,
- the decision to smoke,...

It also raises many novel issues.

In particular, appropriate models are frequently nonlinear.

Regression with Binary Outcomes

Let's first discuss this in the abstract. Say we have the model:

$$y = \beta_0 + \beta_1 x + u$$

and we assume that assumptions SLR.1-4 hold.

A question of interpretation:

What does the fitted value $\hat{y}$ from this regression estimate?

(A) $E(y|x)$.
(B) $\Pr(y = 1|x)$.
(C) Both.
(D) Neither.

Regression with Binary Outcomes

Let's first discuss this in the abstract. Say we have the model:

$$y = \beta_0 + \beta_1 x + u$$

and we assume that assumptions SLR.1-4 hold.

A question of interpretation:

What does the fitted value $\hat{y}$ from this regression estimate?

(A) $E(y|x)$.
(B) $\Pr(y = 1|x)$.
(C) Both.
(D) Neither.

Answer:

Both. Recall that if y can only take value 0 or 1, then $E(y) = \Pr(y = 1)$.

But this prediction is:

- typically a value that y cannot actually take,
- sometimes not even in $[0, 1]$.

Regression with Binary Outcomes

Let's first discuss this in the abstract. Say we have the model:

$$y = \beta_0 + \beta_1 x + u$$

and we assume that assumptions **SLR.1-4** hold.

First issue: The predicted y may be neither 0 nor 1

In economic applications, this is typically not an issue.
For understanding how x affects the expected value of y, it does not matter.
It is also fine for prediction under square loss.

Regression with Binary Outcomes

Let's first discuss this in the abstract. Say we have the model:

$$y = \beta_0 + \beta_1 x + u$$

and we assume that assumptions **SLR.1-4** hold.

Second issue: The predicted y may be outside $[0, 1]$

This is a real issue because the fitted value is now not interpretable.

It is a major reason why nonlinear models for this setting were developed.

The above model (which is called **Linear Probability Model**) should not be used if the issue is pervasive in the data at hand.

Regression with Binary Outcomes

The Linear Probability Model

$$y = \beta_0 + \beta_1 x + u$$

We said we assume **SLR.1-4**.

However, note these are quite restrictive.

They imply that $\Pr(y = 1|x) = \beta_0 + \beta_1 x$.
If $\beta_1 \neq 0$, this is only logically possible if the possible values of x are retsricted.

In this same case of $\beta_1 \neq 0$, what about assumptions **SLR.5** (homoskedasticity) and **SLR.6** (normality)?

(A) Homoskedasticity may or may not hold.
(B) Homoskedasticity cannot possibly hold.

Regression with Binary Outcomes

The Linear Probability Model

$$y = \beta_0 + \beta_1 x + u$$

We said we assume **SLR.1-4**.

However, note these are quite restrictive.

They imply that $\Pr(y = 1|x) = \beta_0 + \beta_1 x$.
If $\beta_1 \neq 0$, this is only logically possible if the possible values of x are retsricted.

In this same case of $\beta_1 \neq 0$, what about assumptions **SLR.5** (homoskedasticity) and **SLR.6** (normality)?

(A) Homoskedasticity may or may not hold.
(B) Homoskedasticity cannot possibly hold.

Homoskedasticity cannot hold:
As $(y|x)$ is binary with $\Pr(y = 1|x) = \beta_0 + \beta_1 x$,

$$\text{Var}(y|x) = (\beta_0 + \beta_1 x)(1 - \beta_0 - \beta_1 x).$$

Regression with Binary Outcomes

The Linear Probability Model

$$y = \beta_0 + \beta_1 x + u$$

We said we assume **SLR.1-4**.

However, note these are quite restrictive.

They imply that $\Pr(y = 1|x) = \beta_0 + \beta_1 x$.
If $\beta_1 \neq 0$, this is only logically possible if the possible values of x are retsricted.

In this same case of $\beta_1 \neq 0$, what about assumptions **SLR.5** (homoskedasticity) and **SLR.6** (normality)?

(A) Homoskedasticity may or may not hold.
(B) Homoskedasticity cannot possibly hold.

Homoskedasticity cannot hold:
As $(y|x)$ is binary with $\Pr(y = 1|x) = \beta_0 + \beta_1 x$,

$$\text{Var}(y|x) = (\beta_0 + \beta_1 x)(1 - \beta_0 - \beta_1 x).$$

(A) Normality may or may not hold.
(B) Normality cannot possibly hold.

Regression with Binary Outcomes

The Linear Probability Model

$$y = \beta_0 + \beta_1 x + u$$

We said we assume **SLR.1-4**.

However, note these are quite restrictive.

They imply that $\Pr(y = 1|x) = \beta_0 + \beta_1 x$.
If $\beta_1 \neq 0$, this is only logically possible if the possible values of x are retsricted.

In this same case of $\beta_1 \neq 0$, what about assumptions **SLR.5** (homoskedasticity) and **SLR.6** (normality)?

(A) Homoskedasticity may or may not hold.
(B) Homoskedasticity cannot possibly hold.

Homoskedasticity cannot hold:
As $(y|x)$ is binary with $\Pr(y = 1|x) = \beta_0 + \beta_1 x$,

$$\text{Var}(y|x) = (\beta_0 + \beta_1 x)(1 - \beta_0 - \beta_1 x).$$

(A) Normality may or may not hold.
(B) Normality cannot possibly hold.

Normality cannot possibly hold because y is discrete.

Regression with Binary Outcomes

The Linear Probability Model

Despite these misgivings, the linear probability model is popular in practice and we will now look at an example.

Technically, this is just OLS, the novelty is entirely in the interpretation.

Note the following:

- The model necessarily has heteroskedasticity, so standard errors should reflect that (not always done in practice).
- Since we have an exact model of heteroskedasticty:

$$\text{Var}(y) = E(y)(1 - E(y)),$$

  could consider FGLS, estimating weights from fitted values in a preliminary OLS regression.
  This is rarely done in practice because even one fitted value outside $[0, 1]$ leads to ill-defined weights.
- In "nice" cases where all fitted values are well interior to $[0, 1]$, none of this really matters.

Empirical Example: Female Labor Force Participation

Econometrica, Vol. 55, No. 4 (July, 1987), 765-799

THE SENSITIVITY OF AN EMPIRICAL MODEL OF MARRIED WOMEN'S HOURS OF WORK TO ECONOMIC AND STATISTICAL ASSUMPTIONS

BY THOMAS A. MROZ[1]

This study undertakes a systematic analysis of several theoretic and statistical assumptions used in many empirical models of female labor supply. Using a single data set (PSID 1975 labor supply data) we are able to replicate most of the range of estimated income and substitution effects found in previous studies in this field. We undertake extensive specification tests and find that most of this range should be rejected due to statistical and model misspecifications. The two most important assumptions appear to be (i) the Tobit assumption used to control for self-selection into the labor force and (ii) exogeneity assumptions on the wife's wage rate and her labor market experience. The Tobit models exaggerate both the income and wage effects. The exogeneity assumptions induce an upwards bias in the estimated wage effect; the bias due to the exogeneity assumption on the wife's labor market experience, however, substantially diminishes when one controls for self-selection into the labor force through the use of unrestricted generalized Tobit procedures. An examination of the maintained assumptions in previous studies further supports these results. These inferences suggest that the small responses to variations in wage rates and nonwife income found here provide a more accurate description of the behavioral responses of working married women than those found in most previous studies.

KEYWORDS: Female labor supply, specification tests, sample selection biases, taxes and labor supply.

We will illustrate with data from the above paper.
The original data source is the *Panel Study of Income Dynamics* (PSID).

Regression with Binary Outcomes

Empirical Example: Female Labor Force Participation

```
      Source |       SS           df       MS      Number of obs   =       753
-------------+----------------------------------   F(7, 745)       =     38.22
       Model |  48.8080578         7  6.97257969   Prob > F        =    0.0000
    Residual |  135.919698       745  .182442547   R-squared       =    0.2642
-------------+----------------------------------   Adj R-squared   =    0.2573
       Total |  184.727756       752  .245648611   Root MSE        =    .42713

------------------------------------------------------------------------------
        inlf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nwifeinc |  -.0034052   .0014485    -2.35   0.019    -.0062488   -.0005616
        educ |   .0379953    .007376     5.15   0.000      .023515    .0524756
       exper |   .0394924   .0056727     6.96   0.000     .0283561    .0506287
     expersq |  -.0005963   .0001848    -3.23   0.001    -.0009591   -.0002335
         age |  -.0160908   .0024847    -6.48   0.000    -.0209686    -.011213
     kidslt6 |  -.2618105   .0335058    -7.81   0.000    -.3275875   -.1960335
     kidsge6 |   .0130122    .013196     0.99   0.324    -.0128935    .0389179
       _cons |   .5855192    .154178     3.80   0.000     .2828442    .8881943
------------------------------------------------------------------------------
```

This is plain vanilla MLS.

It successfully replicates a display in the textbook.

Empirical Example: Female Labor Force Participation

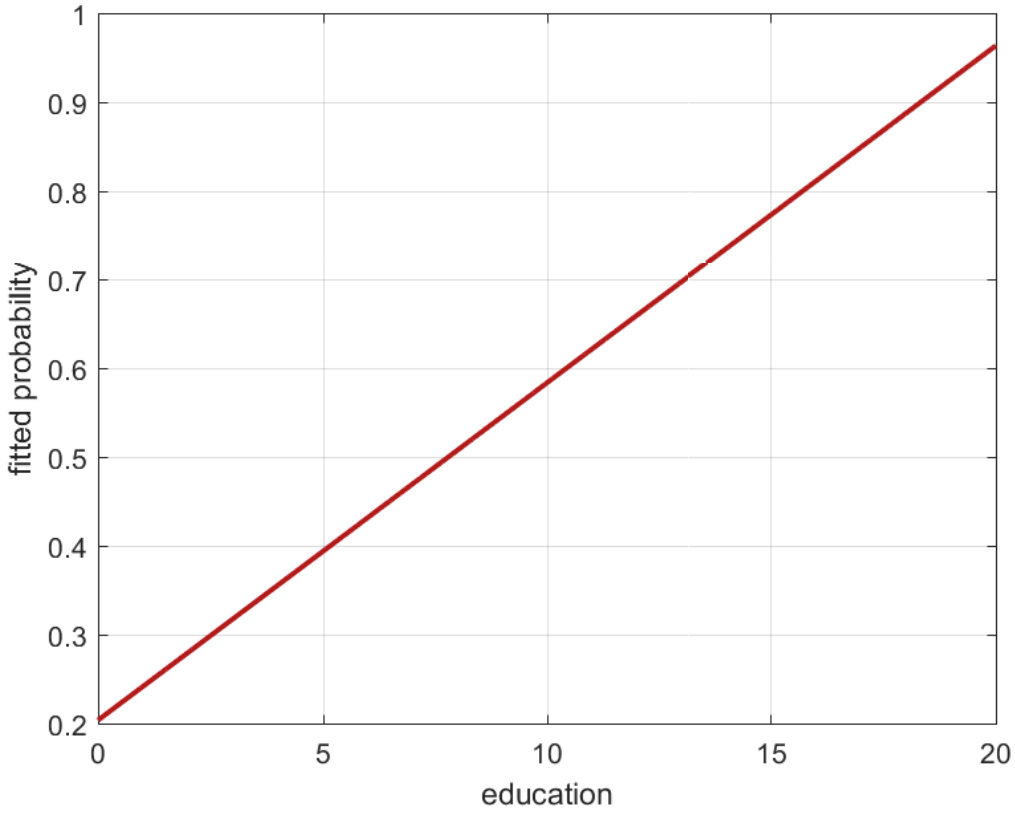
```
Linear regression                               Number of obs     =        753
                                                F(7, 745)         =      62.48
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2642
                                                Root MSE          =     .42713

------------------------------------------------------------------------------
             |               Robust
        inlf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nwifeinc |  -.0034052   .0015249    -2.23   0.026    -.0063988   -.0004115
        educ |   .0379953    .007266     5.23   0.000      .023731    .0522596
       exper |   .0394924     .00581     6.80   0.000     .0280864    .0508983
     expersq |  -.0005963     .00019    -3.14   0.002    -.0009693   -.0002233
         age |  -.0160908    .002399    -6.71   0.000    -.0208004   -.0113812
     kidslt6 |  -.2618105   .0317832    -8.24   0.000    -.3242058   -.1994152
     kidsge6 |   .0130122   .0135329     0.96   0.337     -.013555    .0395795
       _cons |   .5855192   .1522599     3.85   0.000     .2866098    .8844287
------------------------------------------------------------------------------
```

Same with heteroskedasticity robust standard errors.

The difference is practically negligible.

Empirical Example: Female Labor Force Participation



This is the linear fit for *nwifeinc* = 20, *age* = 40, *exper* = 10, and no kids.

The fitted values would become impossible for values of *educ* that are well below respectively above any values observed in the data (the largest one being 17).

In all, the linear probability model is interpretable here.

Regression with Binary Outcomes

Nonlinear Models for Binary Outcomes

The Linear Probability Model is easy to interpret and does a good job in "nice" cases.

Nonetheless, interpretability is a concern and will be first-order if we model small probabilities.
Small probabilities are important:

- probabilities of mortgage or loan default,
- probabilities of specific conditions or diseases,
- "black swans,"...

We therefore look next at models whose fitted values are always in $[0, 1]$.

There are several ways to get there, so we will derive them several times.

Regression with Binary Outcomes

"Pragmatic" Derivation

The basic issue with

$$E(y|x) = \beta_0 + \beta_1 x$$

is that $E(y|x)$ is necessarily in $[0, 1]$ but $\beta_0 + \beta_1 x$ may not be.

Idea:
Let's transform the right-hand side so it is forced to be in $[0, 1]$:

$$E(y|x) = G(\beta_0 + \beta_1 x),$$

where $G(\cdot)$ is a function of our invention that:

- Maps any real number into $[0, 1]$.
- Is strictly increasing.
  (This maintains that $E(y|x)$ is strictly increasing in $\beta_0 + \beta_1 x$ and therefore interpretability of the sign of β_1.)
- Asymptotes 0 for very small arguments and 1 for very large ones.
  (So that in principle, we can model all sorts of probabilities.)

Regression with Binary Outcomes

"Pragmatic" Derivation

The basic issue with

$$E(y|x) = \beta_0 + \beta_1 x$$

is that $E(y|x)$ is necessarily in $[0, 1]$ but $\beta_0 + \beta_1 x$ may not be.

Idea:
Let's transform the right-hand side so it is forced to be in $[0, 1]$:

$$E(y|x) = G(\beta_0 + \beta_1 x),$$

where the function $G(\cdot)$:

- Maps any real number into $[0, 1]$.
- Is strictly increasing.
- Asymptotes 0 for very small arguments and 1 for very large ones.

Which of the following could we **not** use?

(A) The c.d.f. of the standard normal distribution.
(B) The c.d.f. of the uniform distribution on $[0, 1]$.
(C) The c.d.f. of the t_5-distribution.

Regression with Binary Outcomes

"Pragmatic" Derivation

The basic issue with

$$E(y|x) = \beta_0 + \beta_1 x$$

is that $E(y|x)$ is necessarily in $[0, 1]$ but $\beta_0 + \beta_1 x$ may not be.

Idea:
Let's transform the right-hand side so it is forced to be in $[0, 1]$:

$$E(y|x) = G(\beta_0 + \beta_1 x),$$

where the function $G(\cdot)$:

- Maps any real number into $[0, 1]$.
- Is strictly increasing.
- Asymptotes 0 for very small arguments and 1 for very large ones.

We could use any c.d.f. of a random variable that is continuously distributed on (all of) the real line.

Conversely, any function $G(\cdot)$ that we could use can be interpreted as a c.d.f.! (The above properties are sufficient for a function to be a c.d.f.)

Regression with Binary Outcomes

Probit and Logit

The by far most popular models for binary outcomes are:

- The Probit model

$$E(y|x) = \Phi(\beta_0 + \beta_1 x),$$

  where $\Phi(\cdot)$ is the standard normal c.d.f.

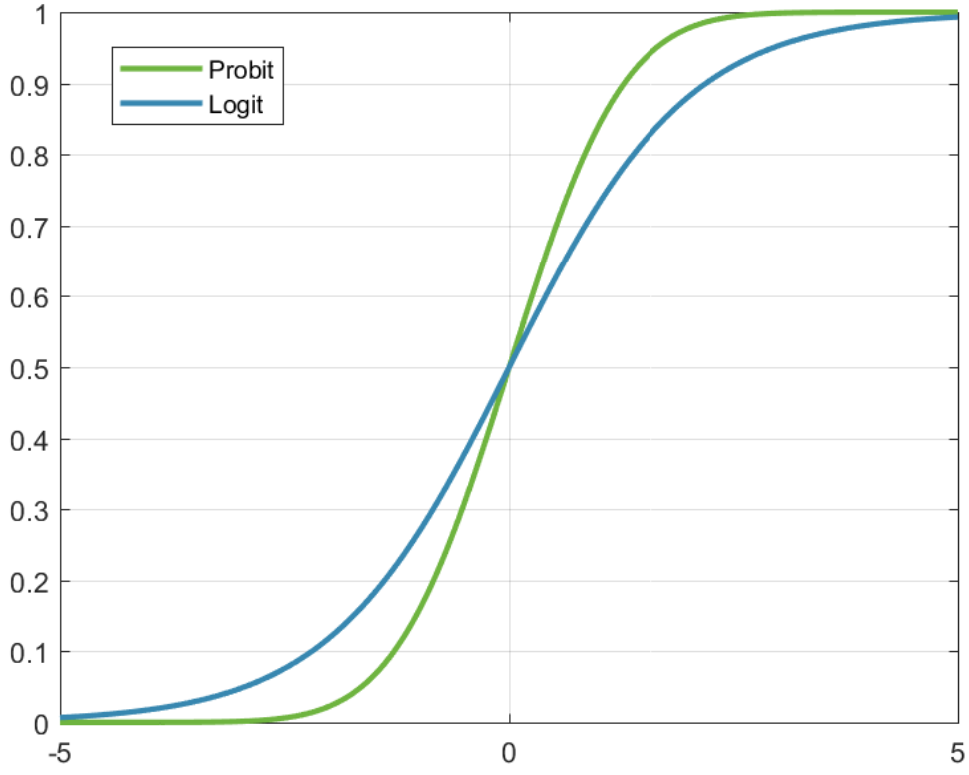
- The Logit model (a.k.a. logistic regression)

$$E(y|x) = \text{logitcdf}(\beta_0 + \beta_1 x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}.$$

  This uses the logistic function (and c.d.f. of the logistic distribution), which the second expression writes out in closed form.

Regression with Binary Outcomes

Probit and Logit

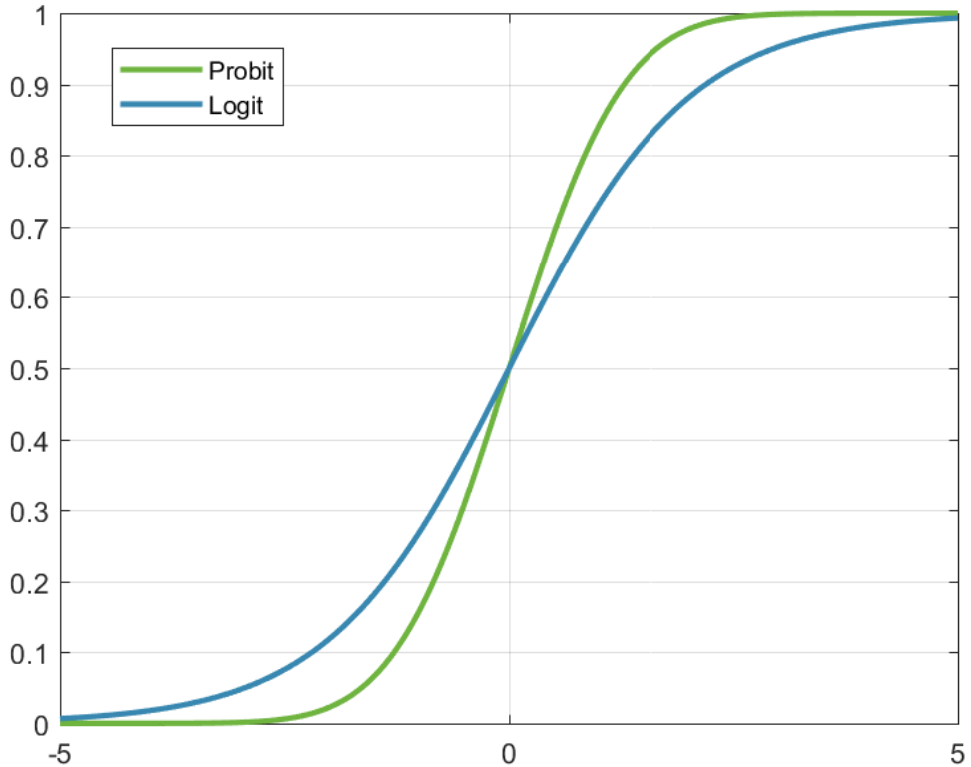


This is how the transformations look. Note the symmetry:

$$\begin{aligned} \Pr(y = 1|\beta_0 + \beta_1 x = 0) &= 1/2 \\ \Pr(y = 1|\beta_0 + \beta_1 x = -t) &= 1 - \Pr(y = 1|\beta_0 + \beta_1 x = t). \end{aligned}$$

Regression with Binary Outcomes

Probit and Logit

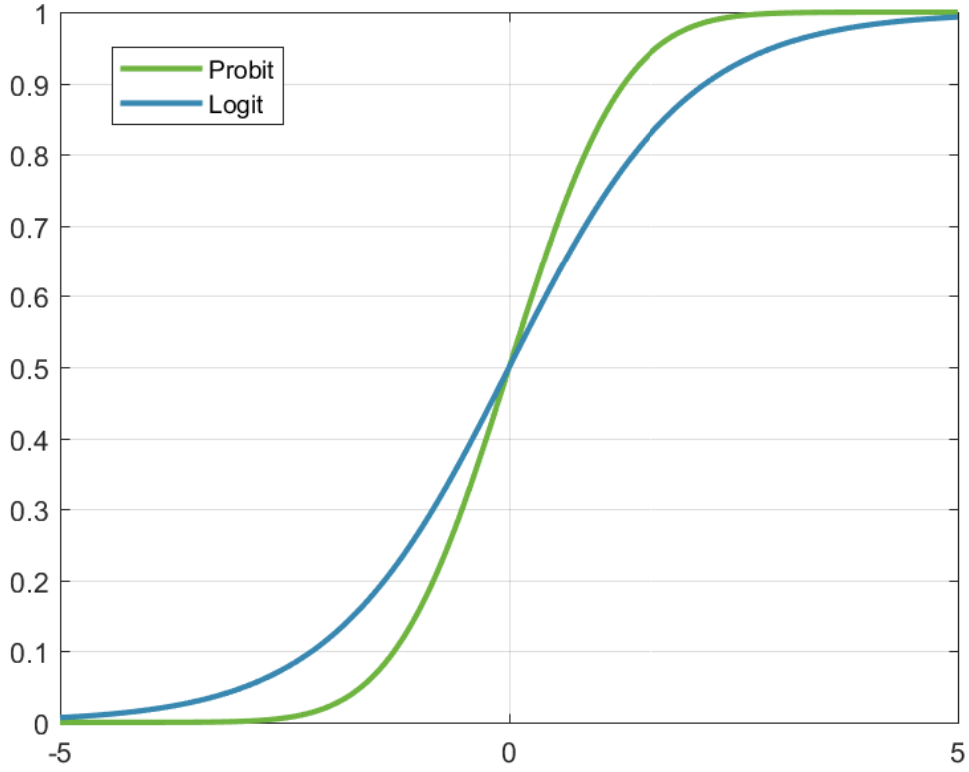


Thus,

- $y = 1$ is more likely than not to occur if $\beta_0 + \beta_1 x > 0$.
- If I flip the sign of the argument, I get the symmetric (about .5) value.

As a result, the *linear index* $\beta_0 + \beta_1 x$ still has some interpretation.

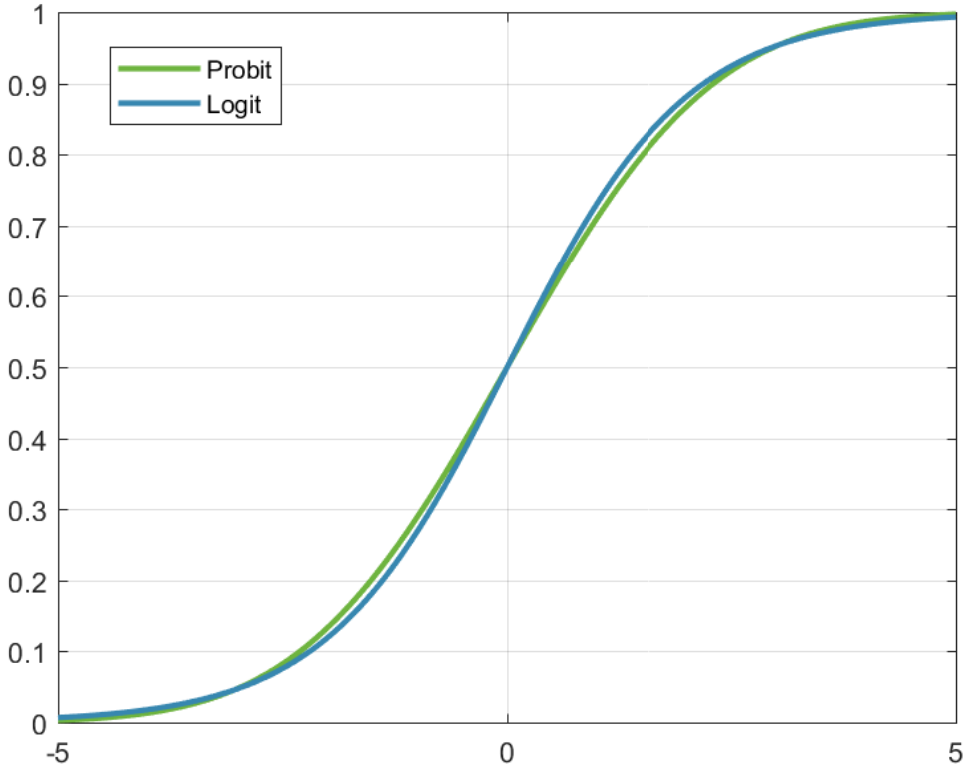
Probit and Logit



How big is the difference?

The transformations as usually defined look rather different.
But we will later learn that the transformation's horizontal scale is arbitrary.
Rescaling the probit,...

Probit and Logit



How big is the difference?

The transformations as usually defined look rather different.
But we will later learn that the transformation's horizontal scale is arbitrary.
Rescaling the probit, we see that the transformations are extremely similar.

Regression with Binary Outcomes

"Economics" Derivation

We can also derive these models as empirical implications of very stylized economics-type models.

Imagine a firm (Trader Joe's, say) enters a given market (Ithaca, say) if expected profits $y^* = \beta_0 + \beta_1 x - u$ are positive.

Here, x collects observables (e.g., size and wealth of a town, proximity to Trader Joe's distribution chain...) and u is unobservable to us but is observed by management.

(Subtracting u is not a typo. It doesn't matter because we get to pick its distribution, but it has a cute effect on the next slide.)

Regression with Binary Outcomes

"Economics" Derivation (ctd.)

Trader Joe's market entry can be modelled as the binary observable y generated as

$$y = \mathbf{1}\{\beta_0 + \beta_1 x - u \geq 0\}$$

and therefore

$$\begin{aligned}
\Pr(y = 1|x) &= \Pr(\mathbf{1}\{\beta_0 + \beta_1 x - u \geq 0\} = 1) \\
&= \Pr(\beta_0 + \beta_1 x - u \geq 0) \\
&= \Pr(u \leq \beta_0 + \beta_1 x) \qquad \leftarrow \text{this is why we subtracted } u \\
&= F(\beta_0 + \beta_1 x),
\end{aligned}$$

where F is the c.d.f. of u. Therefore:

- If we assume that u is standard normal, we get the Probit model.
- If we assume that u is logistic, we get the Logit model.
- And so on.

Regression with Binary Outcomes

"Biostatistics" Derivation of Logit

It can be useful to model causality in terms of relative changes to probabilities.

For example,

- the effect of nutrition on cancer incidence or
- the effect of extant health issues on Covid-19 mortality

might be expressed in terms of the **relative risk** or the **odds ratio**.
(For small probabilities, these are essentially the same.)

Define the **odds** that $y = 1$ as

$$\frac{\Pr(y = 1)}{1 - \Pr(y = 1)} = \text{"odds in favor or against } y = 1\text{"}.$$

Question

The odds will always be an element of

(A) $[0, 1]$
(B) $[0, +\infty)$
(C) $[-1, 1]$
(D) $(-\infty, +\infty)$

Regression with Binary Outcomes

"Biostatistics" Derivation of Logit

It can be useful to model causality in terms of relative changes to probabilities.

For example,

- the effect of nutrition on cancer incidence or
- the effect of extant health issues on Covid-19 mortality

might be expressed in terms of the **relative risk** or the **odds ratio**.
(For small probabilities, these are essentially the same.)

Define the **odds** that $y = 1$ as

$$\frac{\Pr(y = 1)}{1 - \Pr(y = 1)} = \text{"odds in favor or against } y = 1\text{"}.$$

Note:

- The odds are necessarily in $[0, \infty)$.
- Odds of 1 are even odds.
- Odds of and r and $1/r$ are symmetrically opposed:
  They correspond to probabilities π and $1 - \pi$.

"Biostatistics" Derivation of Logit (ctd.)

For a simple model of how x affects odds, we then need our linear index $\beta_0 + \beta_1 x$ to be mapped onto $[0, \infty)$, ideally such that:

- An index of 0 maps onto 1,
- Indices t and $-t$ map onto reciprocals.

There's an easy way to do that:

"Biostatistics" Derivation of Logit (ctd.)

For a simple model of how x affects odds, we then need our linear index $\beta_0 + \beta_1 x$ to be mapped onto $[0, \infty)$, ideally such that:

- An index of 0 maps onto 1,
- Indices t and $-t$ map onto reciprocals.

There's an easy way to do that:
Take the exponential, i.e., map $\beta_0 + \beta_1 x$ onto $\exp(\beta_0 + \beta_1 x)$.

Regression with Binary Outcomes

"Biostatistics" Derivation of Logit (ctd.)

For a simple model of how x affects odds, we then need our linear index $\beta_0 + \beta_1 x$ to be mapped onto $[0, \infty)$, ideally such that:

- An index of 0 maps onto 1,
- Indices t and $-t$ map onto reciprocals.

There's an easy way to do that:
Take the exponential, i.e., map $\beta_0 + \beta_1 x$ onto $\exp(\beta_0 + \beta_1 x)$.

We then have

$$\begin{aligned} \frac{\Pr(y=1|x)}{1-\Pr(y=1|x)} &= \exp(\beta_0 + \beta_1 x) \\ \iff \Pr(y=1|x) &= \frac{\exp(\beta_0 + \beta_1 x)}{1+\exp(\beta_0 + \beta_1 x)}, \end{aligned}$$

i.e. the Logit model.

Regression with Binary Outcomes

Next Steps

Probit and Logit raise some interesting questions:

- How do we interpret the coefficients?
- How do we think of the effect of x on y?
- How do we estimate all this?

We will get there in due course.

To make this more tangible, I'll go back to the empirical illustration first.

For the moment, just assume that I (or Stata, really) know how to estimate this.

Regression with Binary Outcomes

Next Steps

Probit and Logit raise some interesting questions:

- How do we interpret the coefficients?
- How do we think of the effect of x on y?
- How do we estimate all this?

We will get there in due course.

To make this more tangible, I'll go back to the empirical illustration first.

For the moment, just assume that I (or Stata, really) know how to estimate this.

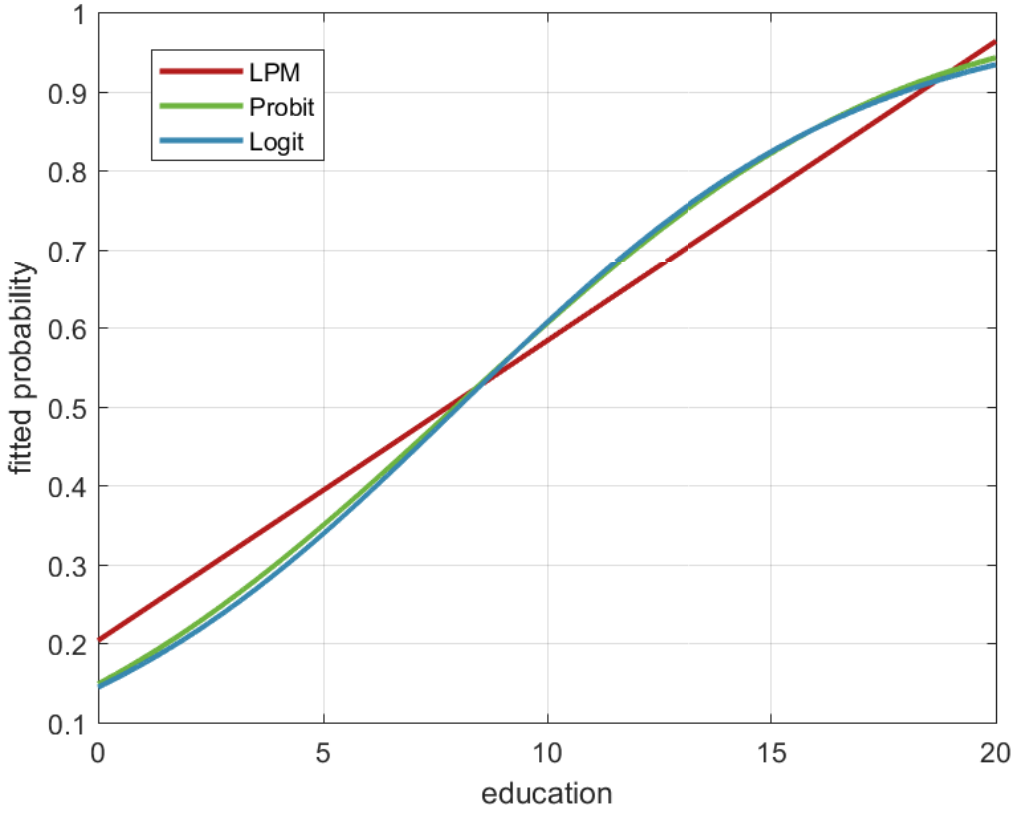
The example is multivariate, so the models really are

$$E(y|x) = \Phi(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) \quad \text{for the probit and}$$

$$E(y|x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)} \quad \text{for the logit.}$$

Regression with Binary Outcomes

Empirical Example: Female Labor Force Participation



The linear fit (again at $age = 40, exper = 10$) is as before.

We can clearly see how logit and probit respect the bounds on $E(y|x)$.

Also, they are somewhat different from the linear fit.

But not from each other.

Regression with Binary Outcomes

Interpreting Coefficients

We now discuss interpretation of the estimates.
Here it will be crucial to look at the multivariate case:

$$E(y|x) = G(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k).$$

In this model, the marginal effect of x_j on $E(y|x)$ is

$$\frac{\partial E(y|x)}{\partial x_j} = \beta_j g(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k),$$

where $g(\cdot)$ is the derivative of $G(\cdot)$.
For example,

$$\begin{aligned} g(t) &= \phi(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) \quad \text{for probit and} \\ g(t) &= \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{(1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k))^2} \quad \text{for logit,} \end{aligned}$$

where $\phi(\cdot)$ is the standard normal p.d.f.

Regression with Binary Outcomes

Interpreting Coefficients

Let's first think about effects of continuous regressors.
Recall:

$$\begin{aligned} E(y|x) &= G(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) \\ \frac{\partial E(y|x)}{\partial x_j} &= \beta_j g(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k). \end{aligned}$$

Question:

Do the individual coefficients β_j still contain some information?

(A) No, because you need to evaluate the whole $\beta_j g(X\beta)$ expression to interpret the effects of x_j.
(B) Yes, because the sign of β_j will tell you weather the change in the x_j makes the outcome more or less likely.
(C) Yes, because we can tell the relative magnitudes of the effects just by looking at β_j's.
(D) Yes, for both reasons B and C.

Regression with Binary Outcomes

Interpreting Coefficients

Let's first think about effects of continuous regressors.
Recall:

$$\begin{aligned} E(y|x) &= G(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) \\ \frac{\partial E(y|x)}{\partial x_j} &= \beta_j g(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k). \end{aligned}$$

The individual coefficients still have some meaning:

- Because $g(\cdot) > 0$, the sign of β_j gives the direction of the effect of x_j.
- Because the factor $g(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$ is the same in all derivatives, the absolute values of coefficients are comparable.

However, the absolute value of one coefficient without context has no meaning.
This is most easily seen by revisiting the "economics" derivation.

Regression with Binary Outcomes

Interpreting Coefficients

Recall the probit model can be interpreted as

$$y = \mathbf{1}\{\beta_0 + \beta_1 x + \cdots + \beta_k x_k - u \geq 0\},$$

where $u \sim N(0, 1)$.

What happens if I double all coefficients and also double u (i.e., assume that $u \sim N(0, 4)$)?

(A) Nothing. The model will remain as before.
(B) The new coefficient estimates will be $1/4$ of what they were before.
(C) The new coefficient estimates will be $1/2$ of what they were before.
(D) The new coefficient estimates will be 2 times what they were before.
(E) The new coefficient estimates will be 4 times what they were before.

Regression with Binary Outcomes

Interpreting Coefficients

Recall the probit model can be interpreted as

$$y = \mathbf{1}\{\beta_0 + \beta_1 x + \cdots + \beta_k x_k - u \geq 0\},$$

where $u \sim N(0, 1)$.

What happens if I double all coefficients and also double u (i.e., assume that $u \sim N(0, 4)$)?

Nothing:

$$\begin{aligned} y &= \mathbf{1}\{2 \times \beta_0 + 2 \times \beta_1 x + \cdots + 2 \times \beta_k x_k - 2 \times u \geq 0\} \\ &= \mathbf{1}\{\beta_0 + \beta_1 x + \cdots + \beta_k x_k - u \geq 0\}, \end{aligned}$$

so the model would be the same.

Regression with Binary Outcomes

Interpreting Coefficients

Recall the probit model can be interpreted as

$$y = \mathbf{1}\{\beta_0 + \beta_1 x + \cdots + \beta_k x_k - u \geq 0\},$$

where $u \sim N(0, 1)$.

We conclude:

- Setting the variance of u to 1 is a normalization.
  (Same for setting u to logistic rather than "twice logistic.")
- Absolute values of individual coefficients are meaningless.
  (Relative magnitudes of different coefficents are not!)
- The absolute value of $g(\cdot)$ at one particular argument is meaningless.
  (The relative magnitude across different x is not!)
- This is why I previously said that the horizontal scaling of the probit transform is arbitrary.

The same observations apply to any choice of $G(\cdot)$.

Interpreting Coefficients

So what is "the effect" of x_j?

As with all nonlinear models, there is no "true" way of expressing this effect in a single number.

But there are some plausible summary measures.
We could report:

- The estimated effect at one or a few values of special interest,
- the estimated effect at the average,
- the average estimated effect.

Regression with Binary Outcomes

Interpreting Coefficients

The estimated effect of x_j at the average is

$$\hat{\beta}_j g(\hat{\beta}_0 + \hat{\beta}_1 \overline{x}_1 + \cdots + \hat{\beta}_k \overline{x}_k).$$

The average estimated effect is

$$\hat{\beta}_j \times \frac{1}{n} \sum_{i=1}^{n} g(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}).$$

These can be numerically quite different!

The estimated effect at the average can be awkward if some regressor other than x_j is discrete.

This can be handled by common-sense adjustment, by switching to the median, or by reporting the average effect.

In the empirical example, I plotted the fit at 0 small kids because I did not want to report predictions for mothers with .27 small kids. The other numbers were rounded sample averages.

Empirical Example: Female Labor Force Participation

```
Probit regression                                 Number of obs     =        753
                                                  LR chi2(7)        =     227.14
                                                  Prob > chi2       =     0.0000
Log likelihood = -401.30219                       Pseudo R2         =     0.2206

------------------------------------------------------------------------------
        inlf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nwifeinc |  -.0120237   .0048398    -2.48   0.013    -.0215096   -.0025378
        educ |   .1309047   .0252542     5.18   0.000     .0814074     .180402
       exper |   .1233476   .0187164     6.59   0.000     .0866641    .1600311
     expersq |  -.0018871      .0006    -3.15   0.002     -.003063   -.0007111
         age |  -.0528527   .0084772    -6.23   0.000    -.0694678   -.0362376
     kidslt6 |  -.8683285   .1185223    -7.33   0.000    -1.100628    -.636029
     kidsge6 |    .036005   .0434768     0.83   0.408     -.049208    .1212179
       _cons |   .2700768    .508593     0.53   0.595    -.7267473    1.266901
------------------------------------------------------------------------------
```

Here is the probit estimation of our empirical example...

Regression with Binary Outcomes

Empirical Example: Female Labor Force Participation

```
Logistic regression                               Number of obs     =        753
                                                  LR chi2(7)        =     226.22
                                                  Prob > chi2       =     0.0000
Log likelihood = -401.76515                       Pseudo R2         =     0.2197

------------------------------------------------------------------------------
        inlf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nwifeinc |  -.0213452   .0084214    -2.53   0.011    -.0378509   -.0048394
        educ |   .2211704   .0434396     5.09   0.000     .1360303    .3063105
       exper |   .2058695   .0320569     6.42   0.000     .1430391    .2686999
     expersq |  -.0031541   .0010161    -3.10   0.002    -.0051456   -.0011626
         age |  -.0880244    .014573    -6.04   0.000     -.116587   -.0594618
     kidslt6 |  -1.443354   .2035849    -7.09   0.000    -1.842373   -1.044335
     kidsge6 |   .0601122   .0747897     0.80   0.422     -.086473    .2066974
       _cons |   .4254524   .8603697     0.49   0.621    -1.260841    2.111746
------------------------------------------------------------------------------
```

...and here is the logit.

The absolute values of coefficients differ by a factor of slightly less than 2.

But their signs agree, and relative magnitudes are very similar.

Regression with Binary Outcomes

Empirical Example: Female Labor Force Participation

	LPM	Probit	Logit
average effect	.0380	.0394	.0395
effect at average	.0380	.0495	.0513
effect at $(20, 12, 10, 40, 0, 0)$	.0380	.0454	.0459
effect at $(20, 16, 10, 40, 0, 0)$	.0380	.0300	.0276
effect at $(10, 10, 5, 35, 0, 0)$	.0380	.0515	.0541

Probit and logit pretty much agree, but there are differences to the linear fit.

In short, probit and logit "price in" saturation effects when estimating the effect at a point where the probability is already high.

The effect at the average is hard to interpret as the average includes fractions of kids, but mathematically it is well-defined.

How to interpret discrete regressors

Conceptually, remarks on discrete regressors are as before.

Note that the effect of a discrete regressor is computed as difference in fitted values as the regressor changes.
This does not in general equal the derivative at either the "before" or "after" value.

Regression with Binary Outcomes

How to interpret discrete regressors

Conceptually, remarks on discrete regressors are as before.

Note that the effect of a discrete regressor is computed as difference in fitted values as the regressor changes.
This does not in general equal the derivative at either the "before" or "after" value.

Confession:
I cheated in the example.
Education is actually discrete.

So let's compute the table again but with the difference of fitted values as one year is *added* to the baseline value.
(I omit the averages. I really don't know what it means to move from 12.2 to 13.2 years of education.)

Regression with Binary Outcomes

Empirical Example: Female Labor Force Participation

	LPM	Probit	Logit
effect at $(20, 12, 10, 40, 0, 0)$	.0380	.0437	.0437
previously:	.0380	.0454	.0459
effect at $(20, 16, 10, 40, 0, 0)$	.0380	.0280	.0256
previously:	.0380	.0300	.0276
effect at $(10, 10, 5, 35, 0, 0)$	.0380	.0507	.0530
previously:	.0380	.0515	.0541

The numbers are similar.

But the saturation effect at $educ = 16$ is even more pronounced.

Regression with Binary Outcomes

How to estimate Logit or Probit

These models are the first *genuinely nonlinear* statistical models in this lecture.

We will informally talk about how to estimate them.

This allows me to briefly mention several estimation principles beyond Least Squares.

In the end, we will take on faith that estimation and inference "work."

Regression with Binary Outcomes

Idea 1: Nonlinear Least Squares

With Ordinary Least Squares, the model implication

$$E(y|x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

motivated the estimator

$$(\hat{\beta}_0, \ldots, \hat{\beta}_k) = \arg \min_{b_0,\ldots,b_k} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - \cdots - b_k x_{ki})^2.$$

Analogously,

$$E(y|x) = G(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$$

motivates

$$(\hat{\beta}_0, \ldots, \hat{\beta}_k) = \arg \min_{b_0,\ldots,b_k} \sum_{i=1}^{n} (y_i - G(b_0 - b_1 x_{1i} - \cdots - b_k x_{ki}))^2.$$

This is the **Nonlinear Least Squares** (NLS) estimator.
It could be used but is not the standard approach.

Regression with Binary Outcomes

Idea 2: Maximum Likelihood

Every observation of $(y, x_1, \ldots, x_k)$ is characterized by

$$\Pr(y = 1|x_1, \ldots, x_k) = G(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k).$$

So for hypothetical values $\tilde{y} \in \{0, 1\}$, we have

$$\begin{aligned} & \Pr(y = \tilde{y}|x_1, \ldots, x_k) \\ = \; & \tilde{y} G(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) + (1 - \tilde{y})(1 - G(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)) \\ = \; & G(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)^{\tilde{y}} \times (1 - G(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k))^{1-\tilde{y}}. \end{aligned}$$

Regression with Binary Outcomes

Idea 2: Maximum Likelihood

Every observation of $(y, x_1, \ldots, x_k)$ is characterized by

$$\Pr(y = 1|x_1, \ldots, x_k) = G(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k).$$

So for hypothetical values $\tilde{y} \in \{0, 1\}$, we have

$$\begin{aligned} & \Pr(y = \tilde{y}|x_1, \ldots, x_k) \\ = \; & \tilde{y} G(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) + (1 - \tilde{y})(1 - G(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)) \\ = \; & G(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)^{\tilde{y}} \times (1 - G(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k))^{1-\tilde{y}}. \end{aligned}$$

Assuming i.i.d. data, the probability of observing any specific, fictitious sequence $(\tilde{y}_1, \ldots, \tilde{y}_n) \in \{0, 1\}^n$, conditionally on all observed covariates, is

$$\begin{aligned} & \Pr((y_1, \ldots, y_n) = (\tilde{y}_1, \ldots, \tilde{y}_n)|x_{11}, \ldots, x_{kn}) \\ = \; & \prod_{i=1}^{n} G(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki})^{\tilde{y}_i} (1 - G(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}))^{1-\tilde{y}_i}. \end{aligned}$$

Regression with Binary Outcomes

Idea 2: Maximum Likelihood

Idea:
Our guess of $(\beta_0, \ldots, \beta_k)$ given the data is whatever maximizes the ex-ante predictive probability of seeing the outcomes that we in fact saw, given covariates.

Formally, the **Maximum Likelihood** estimator $(\hat{\beta}_0, \ldots, \hat{\beta}_k)$ maximizes

$$\prod_{i=1}^{n} G(b_0 + \beta_1 x_{1i} + \cdots + b_k x_{ki})^{y_i} (1 - G(b_0 + b_1 x_{1i} + \cdots + b_k x_{ki}))^{1-y_i}$$

in $(b_0, \ldots, b_k)$, taking the observed values $(y_1, x_{11}, \ldots, x_{kn})$ for granted.

Regression with Binary Outcomes

Idea 2: Maximum Likelihood

Idea:
Our guess of $(\beta_0, \ldots, \beta_k)$ given the data is whatever maximizes the ex-ante predictive probability of seeing the outcomes that we in fact saw, given covariates.

Formally, the **Maximum Likelihood** estimator $(\hat{\beta}_0, \ldots, \hat{\beta}_k)$ maximizes

$$\prod_{i=1}^{n} G(b_0 + \beta_1 x_{1i} + \cdots + b_k x_{ki})^{y_i} (1 - G(b_0 + b_1 x_{1i} + \cdots + b_k x_{ki}))^{1-y_i}$$

in $(b_0, \ldots, b_k)$, taking the observed values $(y_1, x_{11}, \ldots, x_{kn})$ for granted.

Note that this is the same as maximizing the **log likelihood**

$$\sum_{i=1}^{n} [y_i \log G(\cdot) + (1 - y_i) \log(1 - G(\cdot))]$$

or also

$$\frac{1}{n} \sum_{i=1}^{n} [y_i \log G(\cdot) + (1 - y_i) \log(1 - G(\cdot))] .$$

Regression with Binary Outcomes

Idea 2: Maximum Likelihood

Idea:
Our guess of $(\beta_0, \ldots, \beta_k)$ given the data is whatever maximizes the ex-ante predictive probability of seeing the outcomes that we in fact saw, given covariates.

Formally, the **Maximum Likelihood** estimator $(\hat{\beta}_0, \ldots, \hat{\beta}_k)$ maximizes

$$\prod_{i=1}^{n} G(b_0 + \beta_1 x_{1i} + \cdots + b_k x_{ki})^{y_i} (1 - G(b_0 + b_1 x_{1i} + \cdots + b_k x_{ki}))^{1-y_i}$$

in $(b_0, \ldots, b_k)$, taking the observed values $(y_1, x_{11}, \ldots, x_{kn})$ for granted.

Your computer probably maximizes

$$\sum_{i=1}^{n} \left[y_i \log G(\cdot) + (1 - y_i) \log(1 - G(\cdot)) \right].$$

The sum is much easier to handle than the product.

Regression with Binary Outcomes

Idea 2: Maximum Likelihood

Idea:
Our guess of $(\beta_0, \ldots, \beta_k)$ given the data is whatever maximizes the ex-ante predictive probability of seeing the outcomes that we in fact saw, given covariates.

Formally, the **Maximum Likelihood** estimator $(\hat{\beta}_0, \ldots, \hat{\beta}_k)$ maximizes

$$\prod_{i=1}^{n} G(b_0 + \beta_1 x_{1i} + \cdots + b_k x_{ki})^{y_i} (1 - G(b_0 + b_1 x_{1i} + \cdots + b_k x_{ki}))^{1-y_i}$$

in $(b_0, \ldots, b_k)$, taking the observed values $(y_1, x_{11}, \ldots, x_{kn})$ for granted.

A statistician will pretend that you maximize

$$\frac{1}{n} \sum_{i=1}^{n} \left[y_i \log G(\cdot) + (1 - y_i) \log(1 - G(\cdot)) \right].$$

The objective is a sample average $\Rightarrow$ it converges to an expectation as $n \to \infty$.
This is the starting point for proving consistency etc., but we will not go there.

Regression with Binary Outcomes

Idea 2: Maximum Likelihood

Here is a one-slide intuition for why Maximum Likelihood "should" work.
It is a "good" intuition in the sense that the formal proof indeed formalizes it.
The sample objective function

$$\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log G(\cdot) + (1-y_i)\log(1-G(\cdot))\right]$$

can be thought of as "estimator" of the true expected log likelihood

$$E\left(y \log G(b_0 + b_1 x_{1i} + \cdots + b_k x_{ki}) + (1-y)\log(1-G(b_0 + \beta_1 x_{1i} + \cdots + b_k x_{ki})\right).$$

The above function of $(b_0, \ldots, b_k)$ is maximized by the true $(\beta_0, \ldots, \beta_k)$.

This is not obvious! It follows from mathematical insights about relative entropy (a.k.a. Kullback-Leibler divergence) that go beyond this lecture.

But if we take it on faith, then there is an intuition for consistency of Maximum Likelihood. It will be visualized on the next slide.

Regression with Binary Outcomes

Regression with Binary Outcomes

Inference for Maximum Likelihood

Precise formal analysis of Maximum Likelihood goes well beyond this lecture.

There are no simple analogs to our finite-sample results (e.g., unbiasedness) from linear regression.

It is possible to show that, under reasonable assumptions,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N\left(\mathbf{0}, \left(E\left(\frac{g(\boldsymbol{x}^\top\boldsymbol{\beta})^2\boldsymbol{x}\boldsymbol{x}^\top}{G(\boldsymbol{x}^\top\boldsymbol{\beta})(1-G(\boldsymbol{x}^\top\boldsymbol{\beta})}\right)\right)^{-1}\right).$$

The expression resembles earlier ones.

- The denominator is $\Pr(y=1|\boldsymbol{x})(1-\Pr(y=1|\boldsymbol{x}))$.
  Recalling the formula for variance of Bernoulli variables, we recognize it as conditional variance of y. Thus, it is analogous to σ^2.
- The numerator involves $\boldsymbol{x}\boldsymbol{x}^\top$, which is analogous to $\boldsymbol{X}^\top\boldsymbol{X}$ or $(x-E(x))^2$.
  The premultiplying factor adjusts this for the nonlinearity of $\boldsymbol{x}$'s effect on y.
- Numerator and denominator appear flipped, but note the outside $(\cdot)^{-1}$.
  (We cannot just flip the fraction because the numerator is a matrix.)

Regression with Binary Outcomes

Inference for Maximum Likelihood

What do I need to take away from this?

- Asymptotic (not finite sample) inference is possible.
- Under reasonable assumptions, results have similar flavor as before.
- We can use output from estimation packages to test hypotheses and so on.

If you really want to know more:

- ECON 4110,
- Graduate textbooks,
- Grad school!