NATIONAL RESEARCH
UNIVERSITY

Department of Sociology
Laboratory for Comparative Social Research

QUANTITATIVE DATA ANALYSIS

Binary Logistic Regression

Violetta Korsunova

St. Petersburg, 2021

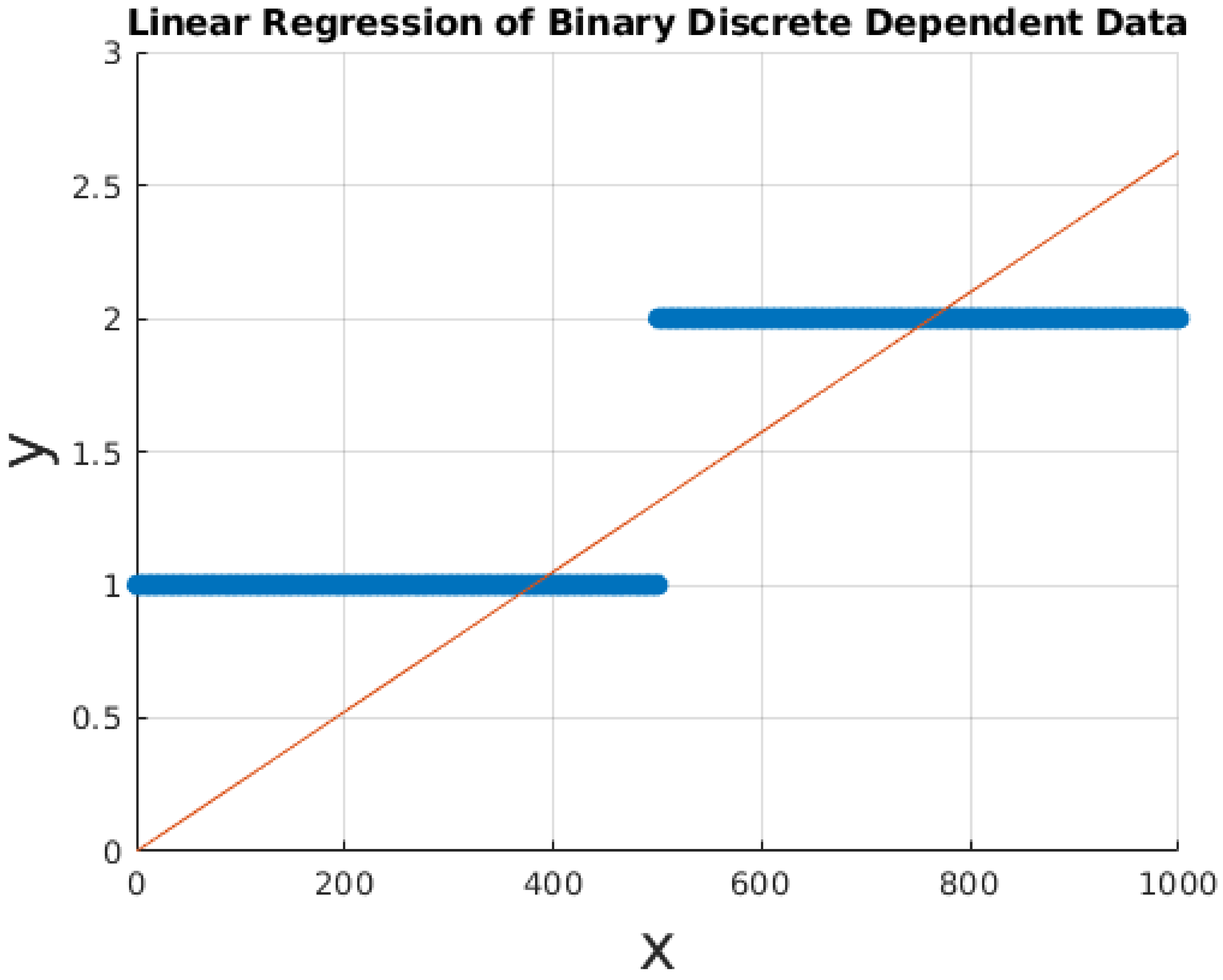
PREVIOUSLY ON QDA

- You already know how to predict the outcomes of continuous variables
- However, social indicators are usually measured with some sort of categorical scales
- Here we are going to discuss how to predict the outcomes of these variables using logistic regression

BINARY (BINOMIAL) VARIABLES

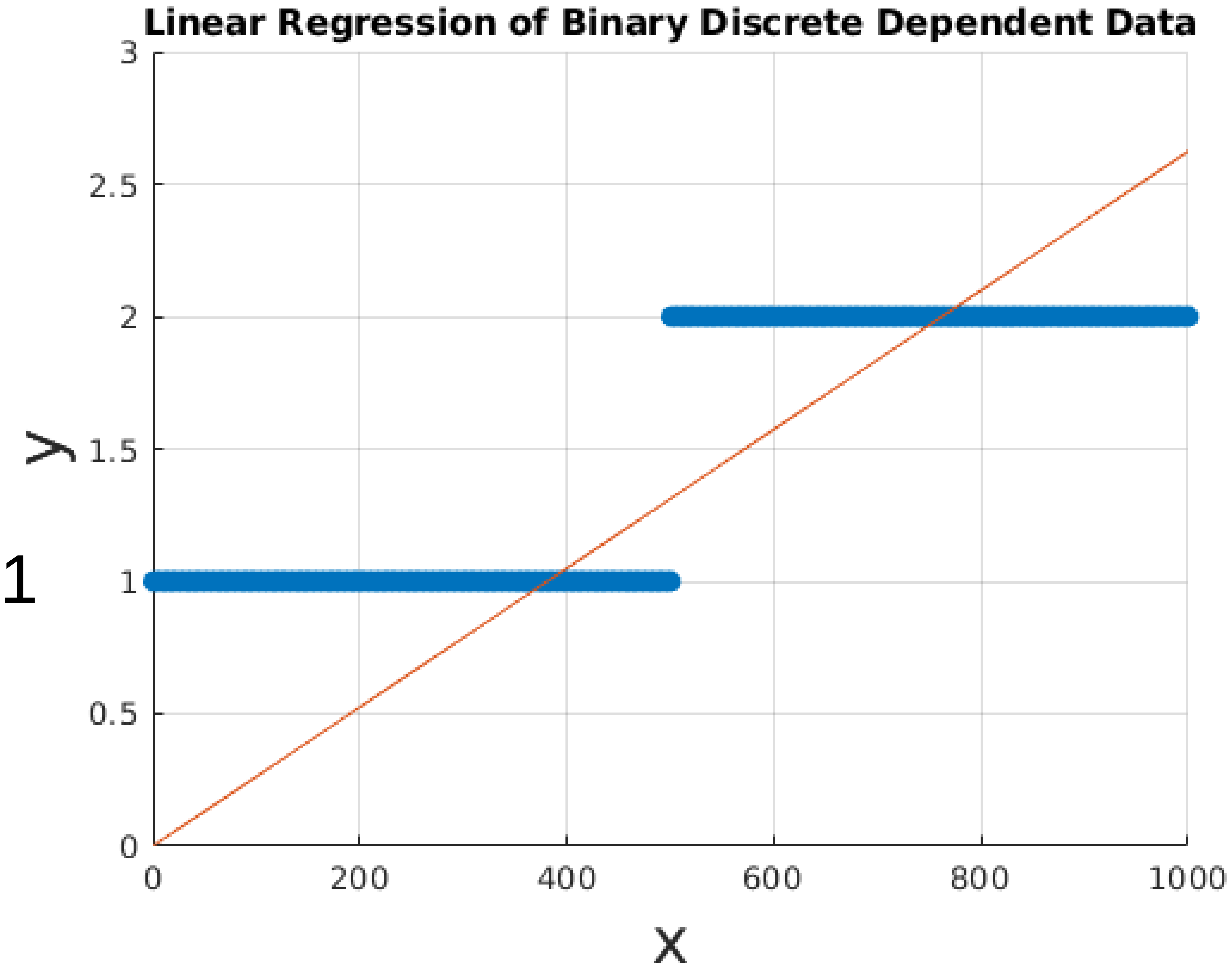
- Variables that only have two possible outcomes: 0, 1; yes, no; A, B, etc.
- The latter outcome is called ‘success’.
- These variables are measured with a binary scale which is discrete and categorical.

WHY NOT JUST USE LINEAR REGRESSION TO PREDICT THESE OUTCOMES?



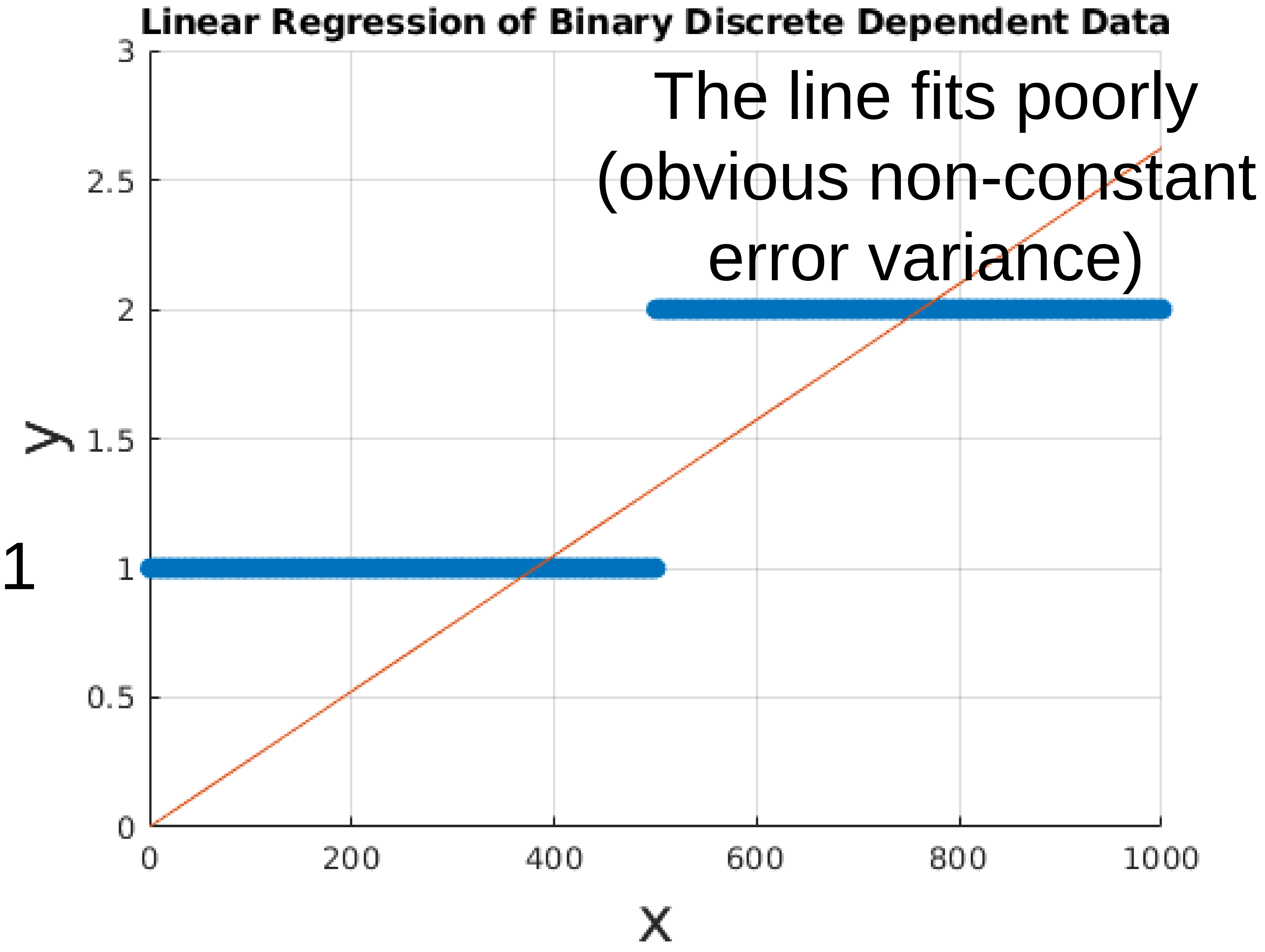
WHY NOT JUST USE LINEAR REGRESSION TO PREDICT THESE OUTCOMES?

The line exceeds 0 and 1

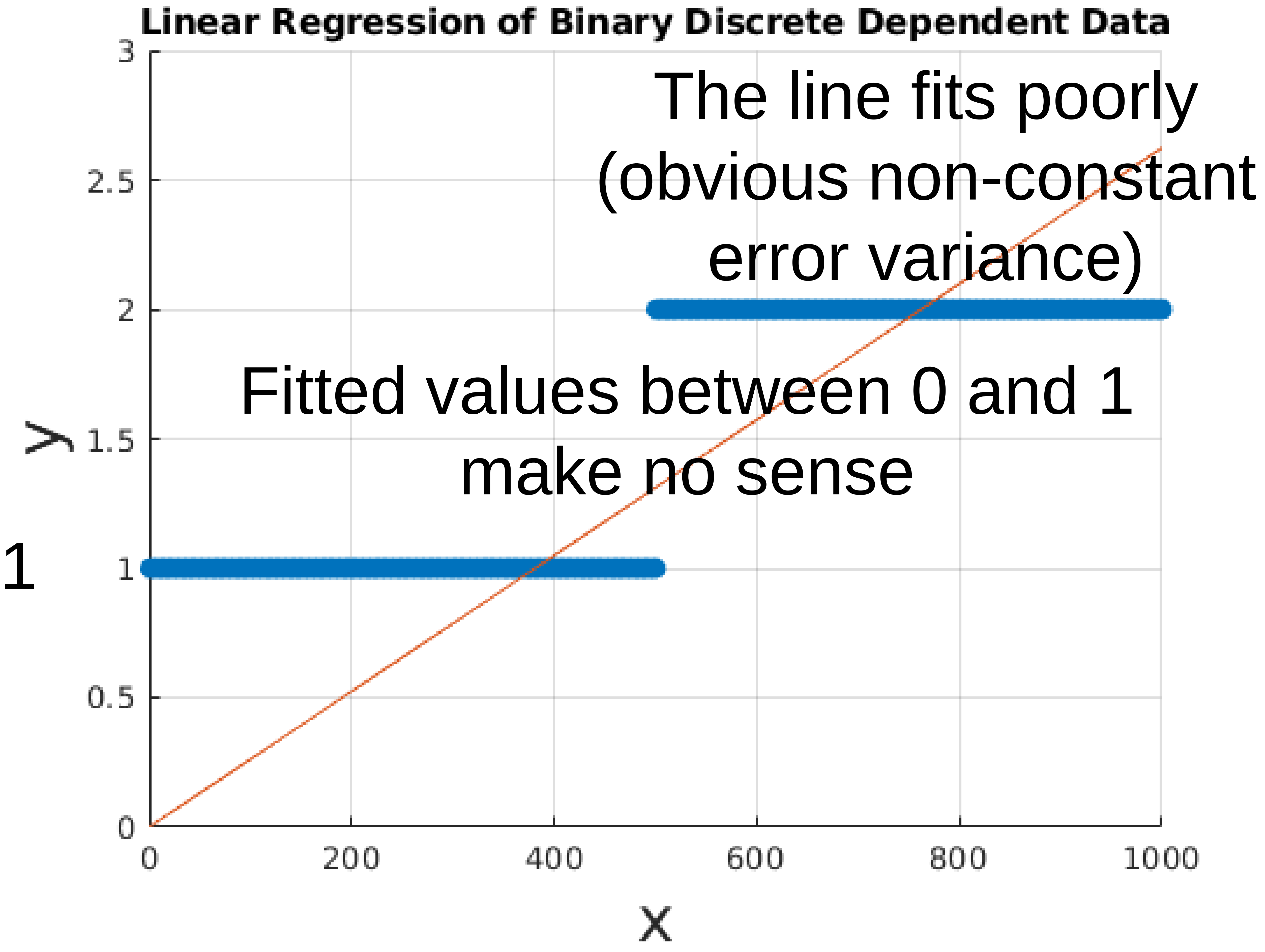


WHY NOT JUST USE LINEAR REGRESSION TO PREDICT THESE OUTCOMES?

The line exceeds 0 and 1



WHY NOT JUST USE LINEAR REGRESSION TO PREDICT THESE OUTCOMES?



The line fits poorly (obvious non-constant error variance)

Fitted values between 0 and 1 make no sense

The line exceeds 0 and 1

SO, WHAT ARE WE GOING TO DO?

- Transform the existing distribution of the outcome variable into something non-discrete and completely defined.
- Although the real outcomes are binomial, the probabilities of outcomes are close to standard normal distribution.
- We can use ‘link-functions’ to link the outcome with probability.

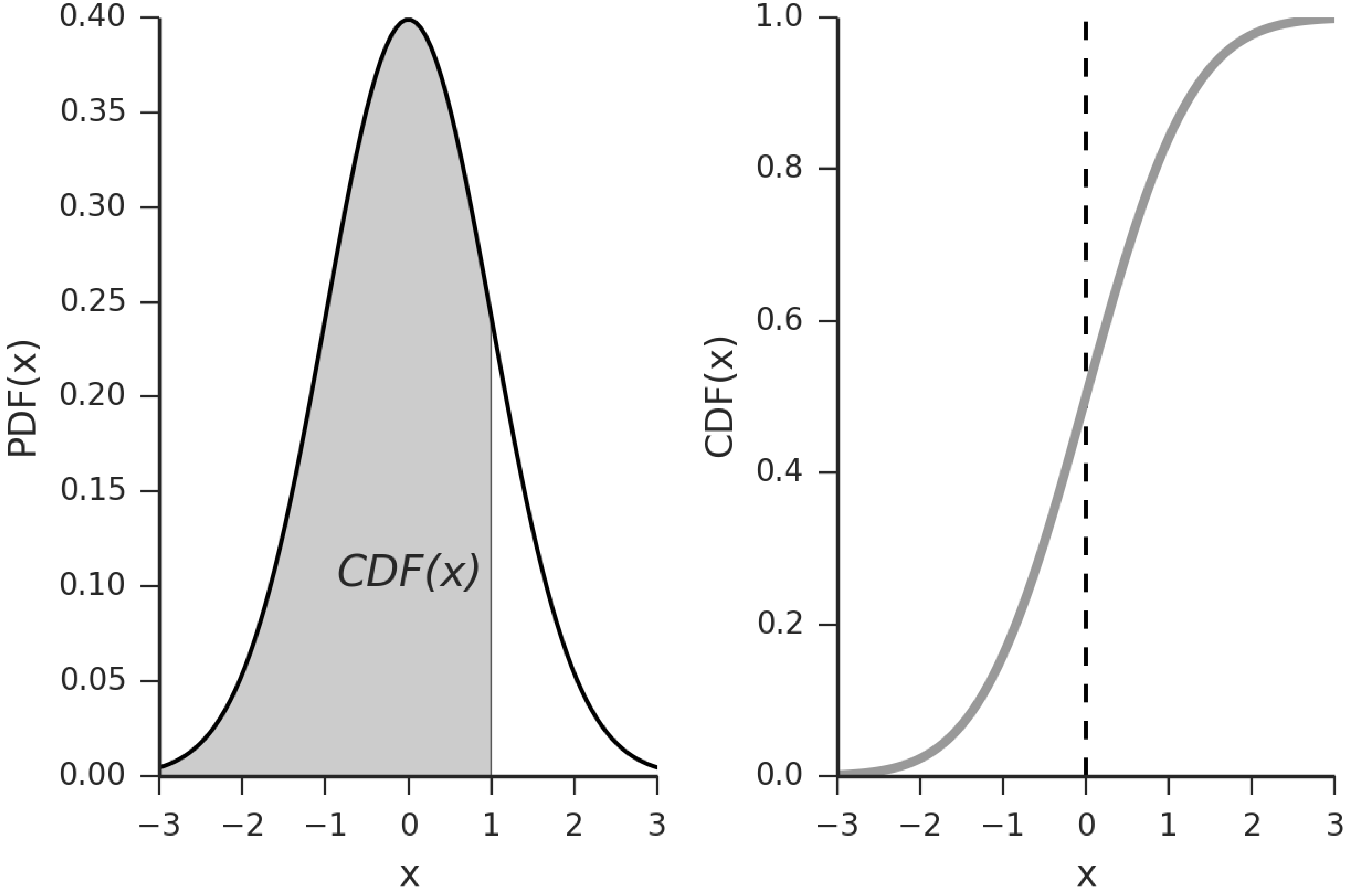
So, let’s go deep down the rabbit hole..

PROBIT LINK FUNCTION

- We assume that there's a normally distributed latent variable Y*:Y* =βX
- However, we only can only observe our outcome variable where:

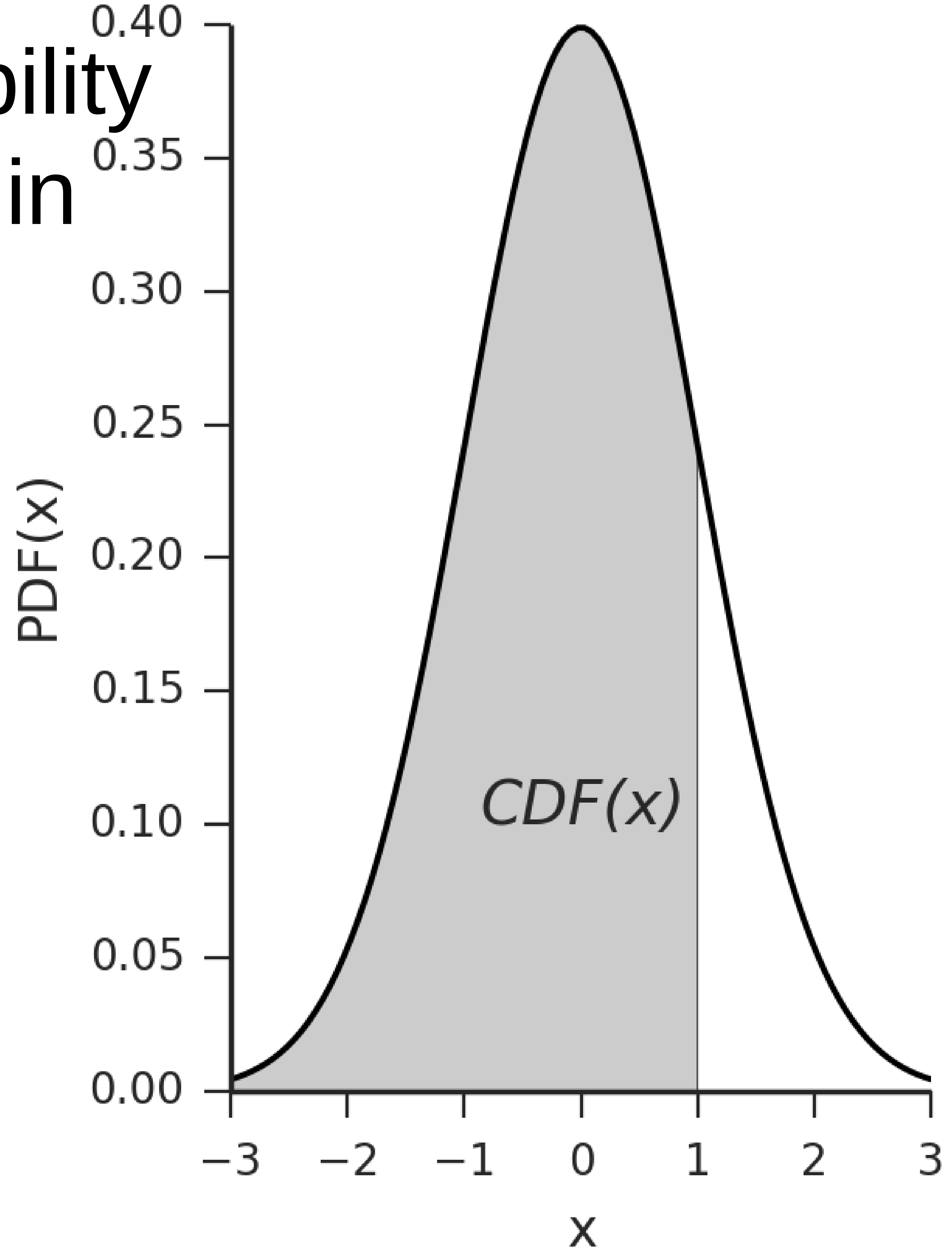
$$\boldsymbol{Y} = \begin{cases} \mathbf{0}\ \boldsymbol{if}\ \boldsymbol{y} * < \boldsymbol{t} \\ \mathbf{1}\ \boldsymbol{if}\ \boldsymbol{y} * \geq \boldsymbol{t} \end{cases}$$

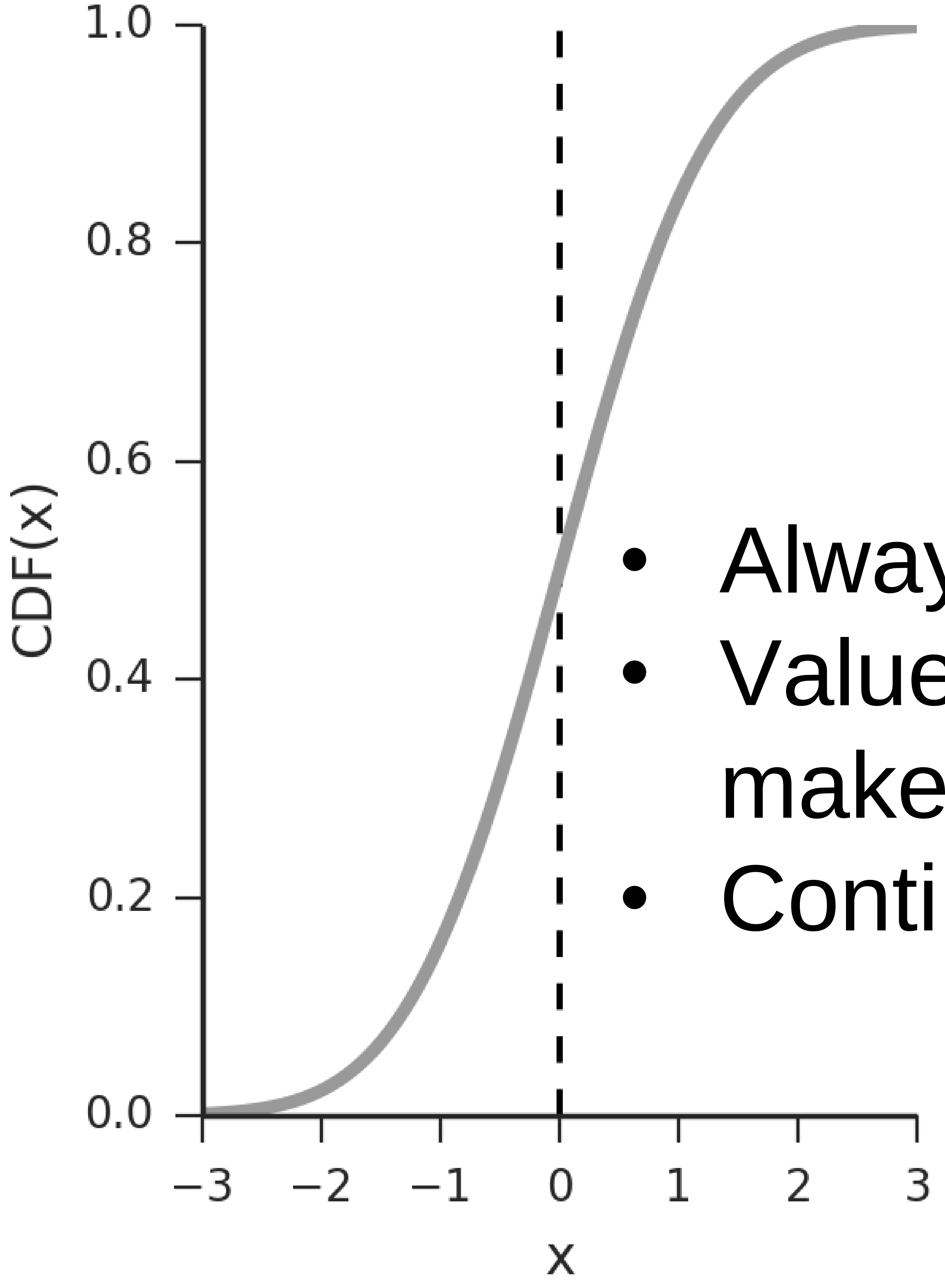
PROBABILITY DISTRIBUTION FUNCTION AND CUMULATIVE DISTRIBUTION FUNCTION



PROBABILITY DISTRIBUTION FUNCTION AND CUMULATIVE DISTRIBUTION FUNCTION

Shows the probability of any outcome in distribution





- Always between 0 and 1
- Values between 0 and 1 make sense
- Continuous

OK, BUT HOW CAN I ESTIMATE THIS?

- In linear regressions we usually use OLS method to find the line that minimizes unexplained variance.
- In logistic regressions we can’t use OLS since no variation is present.
- Instead maximum likelihood estimation is employed

MAXIMUM LIKELIHOOD ESTIMATION

For example, we want to estimate Y=βX:

- Get trial samples of β
- For each β and X calculate Y*

Let's say we got y* = 0.7

Then:

- If y = 1, likelihood is 0.7
- If y = 0, likelihood is 1-0.7 = 0.3

Repeat for each y and set of β

Multiply likelihoods in all samples

Choose the set of β that has the *maximum* likelihood

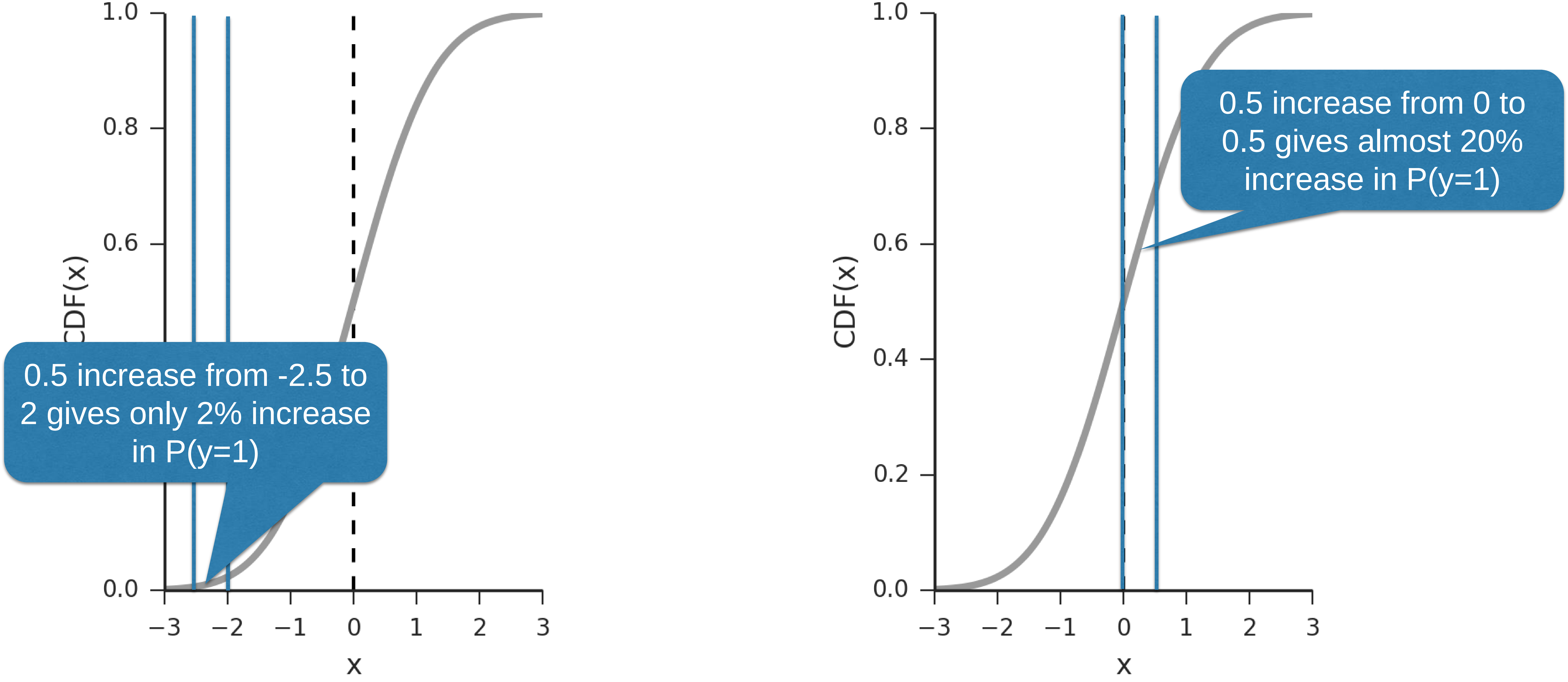
INTERPRETATION OF PROBIT

After you go through estimations, you get β-coefficients and p-values that resemble the results of linear regression, yet, their interpretations are different.
In linear regressions β shows the change in y when x changes by 1.

However, in probit regressions you don't have changes in y, you have *changes in z-scores of probabilities of y=1.*

Moreover, this effect is not constant…

THE EFFECT SIZE DEPENDS ON THE VALUES OF X



MARGINAL EFFECTS

As the effects are volatile, marginal effects need to be presented.

Usually, these are the effects estimated for all mean values of X.

They are interpreted as the average changes in the probability of Y=1 given 1 unit change in X.

LOGIT LINK FUNCTION

Use odds ratio:

$$\frac{P(y = 1)}{P(y = 0)}$$

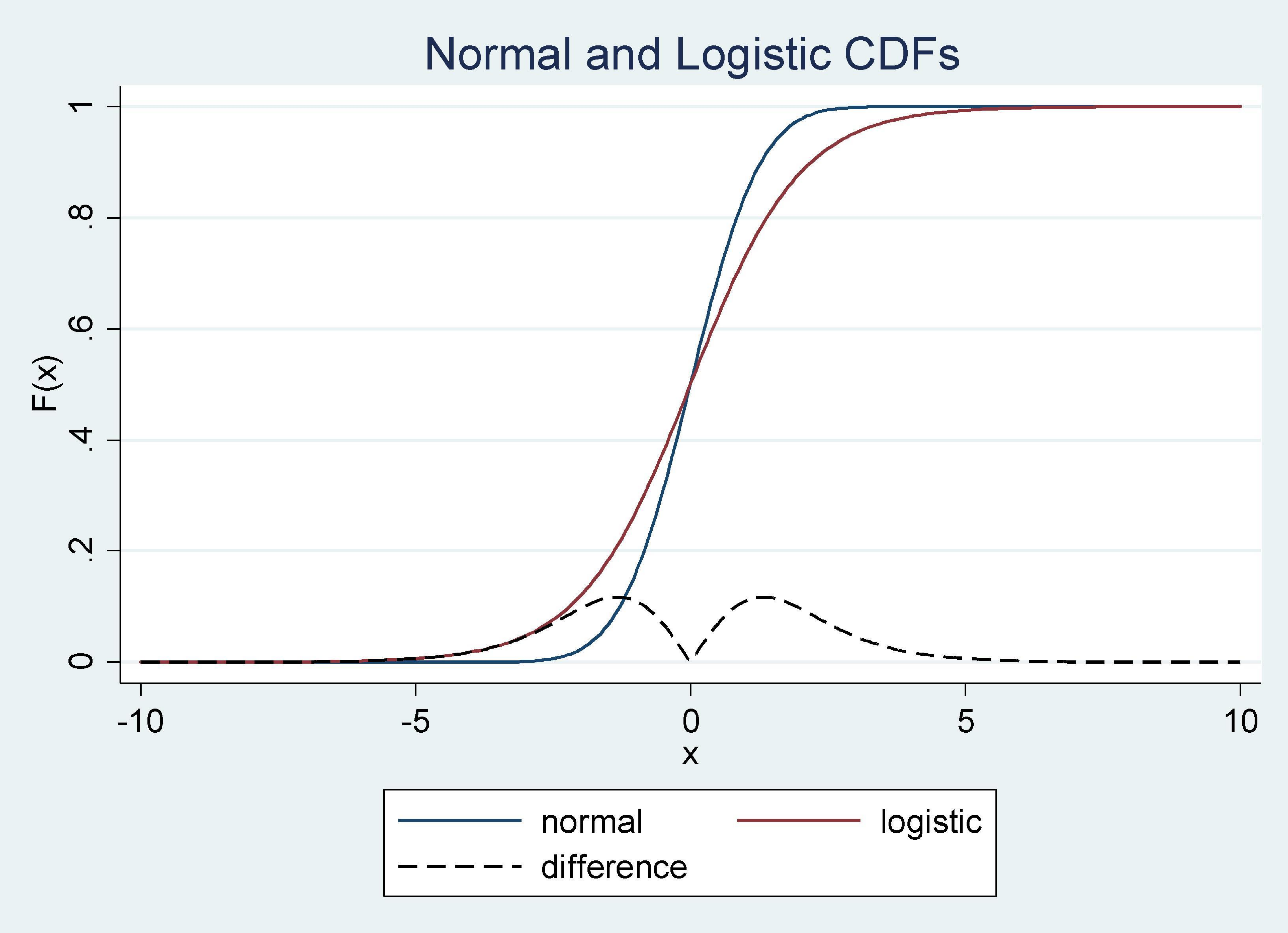
To make this distribution completely defined use the log of odds ratio:

$$\log(\frac{P(y=1)}{P(y=0)})$$

Then link to the linear equation:

$$\log\left(\frac{P(y=1)}{P(y=0)}\right) = \beta X$$

LOGIT AND PROBIT CDFS



INTERPRETATION OF LOGIT

In logit models β shows the *change in log of odds ratio of Y=1* if X changes by 1.

You can calculate exponent to obtain the odds ratio. Then, exp(β) shows the change in odds ratio given change of X by 1.

Don’t forget about marginal effects

PROBIT OR LOGIT?

Generally, both methods provide with similar marginal effects

- Logit is used for true binomial outcomes
- Probit is used when a latent distribution can be hypothesized (e.g. when you split your variable into 2 categories).

NATIONAL RESEARCH
UNIVERSITY