Modeling Binary Outcomes Using Logistic Regression with a Predictor

Part 1 - Introduction to Logistic Regression

Introduction

In many situations, we want to examine how the probability of an event depends on other factors. For example, spending more time studying should increase the probability of earning a passing grade.

Learning Objectives

In this lecture, you will learn the following:

- Modelling the probability (or odds) of an event based on other factors
- Why Ordinary Regression Fails for Binary Data
- Introduction to **Logistic Regression Model**

Example

A random sample of students is selected from a large statistics class.

The following variables are recorded:

- Number of hours studied
- Exam outcome, **pass (P)** or **fail (F)**

Hours	Grade
0	F
0	F
0.5	F
1.5	F
1.5	F
1.5	P
2	F
2.5	F
2.5	F
⋮	⋮
10.5	P
11	P
11	P

*The full dataset '**Hours-and-Grades**' can be downloaded from Brightspace*

Modeling Binary Outcomes Using a Predictor

- In this dataset, we have the **number of hours studied**, which may affect a student's probability of passing the exam.
- So, we can use the number of hours as a predictor to model the probability of passing.
- To do this, let's apply a **regression model**—the same method you learned in your previous statistics class.

$$\underbrace{\boldsymbol{p}}_{\text{Probability of passing}} = A + B * \boldsymbol{Hours}$$

Problems When Using a Regression Model

Problem 1

- To fit a regression model, we need data on **the** **probability of passing** and the number of hours studied for each student.

Probability	?	?	?	?	?	?	?	?	…	?	?	?
Hours	0	0	0.5	1.5	1.5	2	2.5	2.5	…	10.5	11	11

- However, we don't have the data on the **probabilities**!
- Instead, the data we observe—the exam grade—is a **categorical label**: either **Pass** or **Fail**..

Problems When Using a Regression Model

Problem 2

$$\underbrace{p}_{\text{Probability of Passing}} = \underbrace{A + B * Hours}$$

between 0 and 1

$(-\infty, \infty)$

- On the left side, the **probability must be between 0 and 1**
- However, on the right side, a **linear function of the predictor (*e.g.*, hours)** can **produce** values **less than 0 or greater than 1.**
- More importantly, the **linear function** can take **any value** **from $-\infty$ to $+\infty$**.
- In other words, the left side (**probability**) does **not align** with the right side (**linear function**). Therefore, the regression model is **NOT** appropriate.

Logit / log-odd function

One **function** that can perform this mapping is the ***Logit***

$$function\Big(\underbrace{p}_{\text{Probability of passing}}\Big) = Logit(p)$$

$$= \ln\left(\frac{p}{1-p}\right)$$

Nature-**Logarithm**
(simply called ***'log'***)

Odds of an event (e.g. passing)

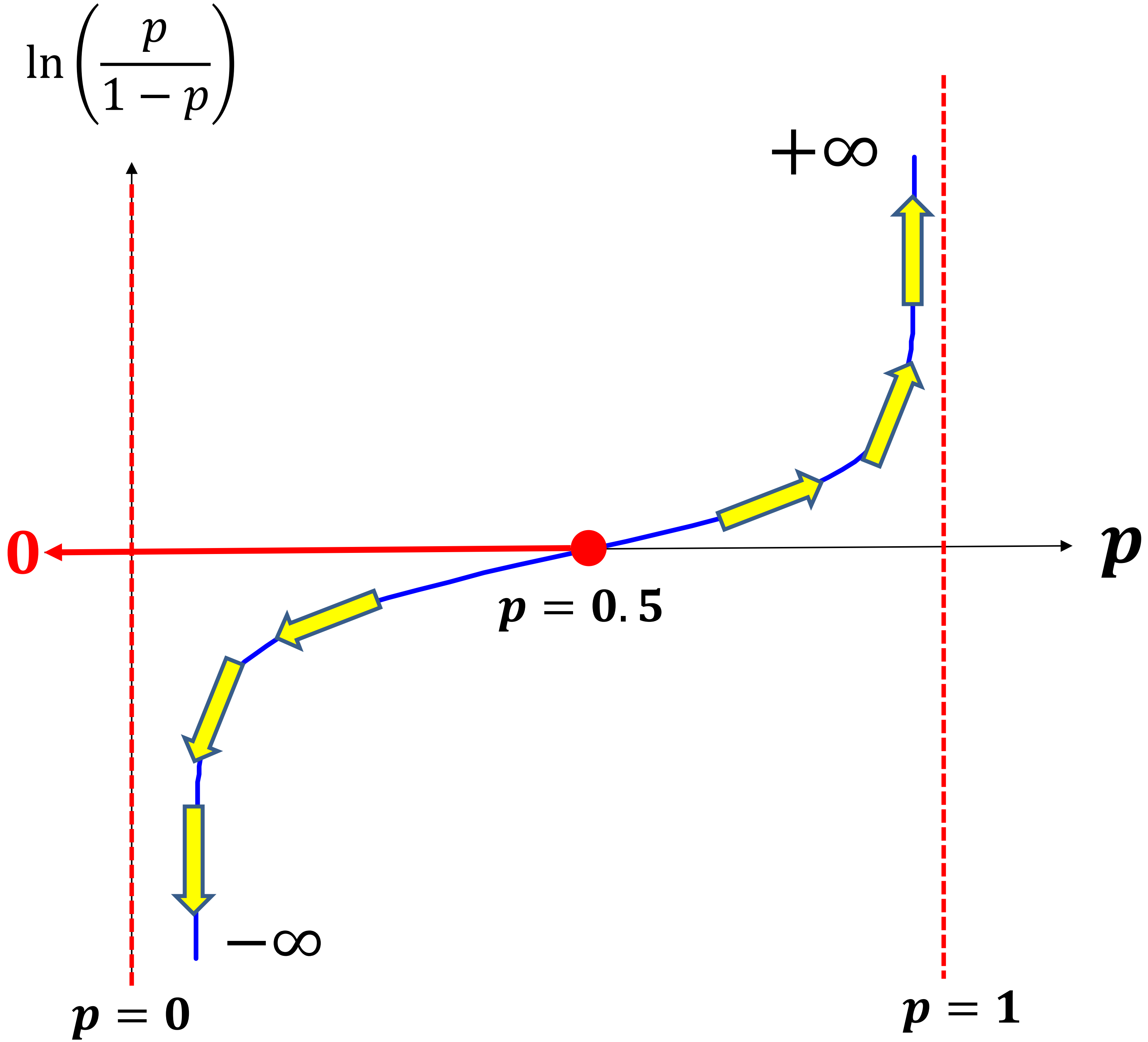
Logit function

$$logit(p) = \underbrace{\ln\left(\frac{p}{1-p}\right)}_{\text{Log-odds}}$$

Let's describe the logit function for p in details

First, we can show that the logit function of p maps any probability between 0 and 1 to a real number in the range $(-\infty, \infty)$

We demonstrate this by graphing the logit function of p



$\ln\left(\frac{p}{1-p}\right)$
$+\infty$
$-\infty$
0
p
$p = 0.5$
$p = 0$
$p = 1$

Let's put every together

$$\underbrace{\ln\left(\frac{p}{1-p}\right)}_{\text{logit function of } p \text{ (log-odds)}} = \underbrace{A + B * x}_{\text{the linear function of the predictor (e.g., hours) as used in a standard regression model}}$$

logit function of p
(log-odds)

the linear function of the predictor (e.g., hours)
as used in a standard regression model

returns any real numbers

$(-\infty, \infty)$

Return any real numbers

$(-\infty, \infty)$

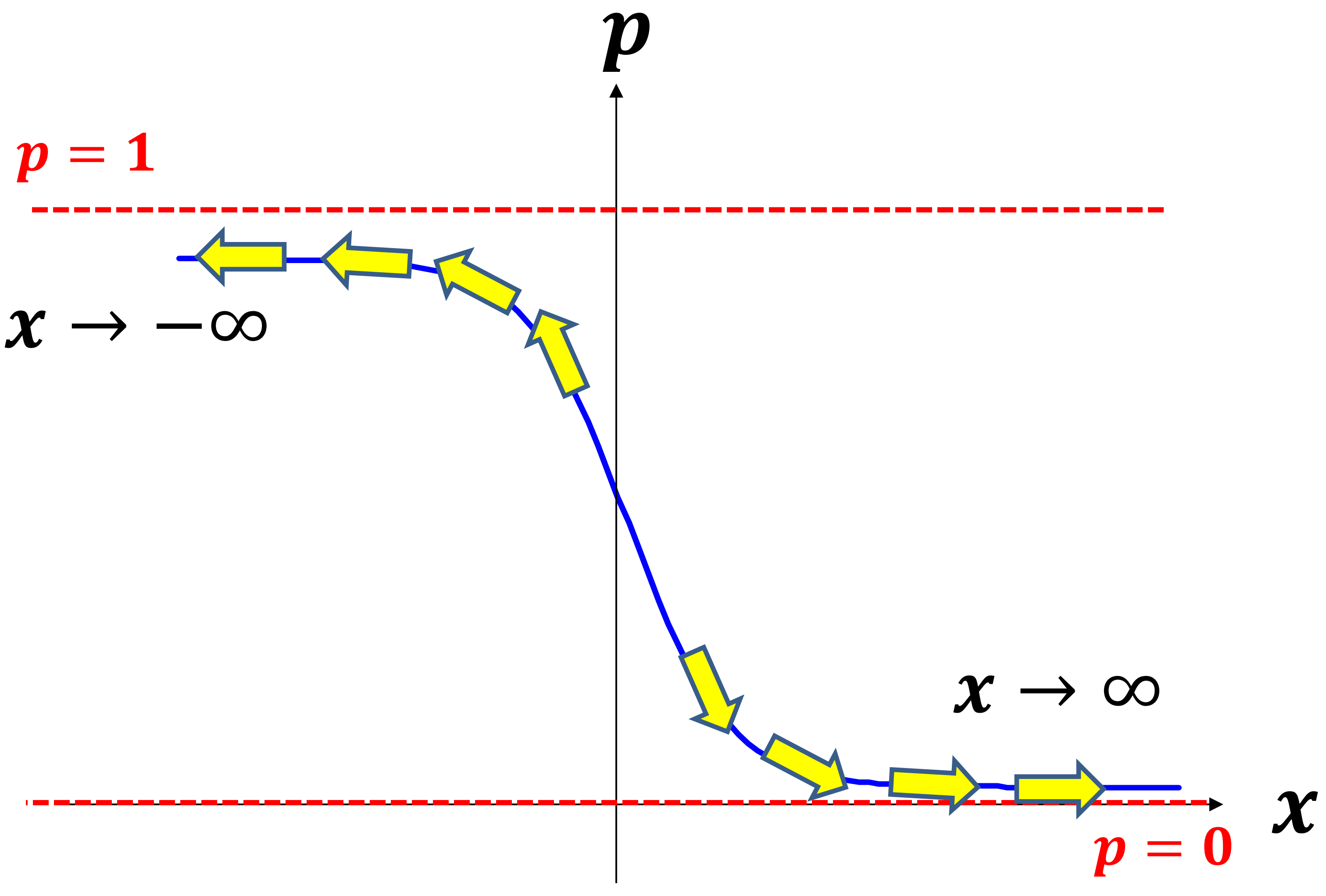
This model is called the **Logistic Regression Model**

(Why is it called '**Logistic**'? I will explain that shortly)

Let's graph the function $p = \frac{e^{A+Bx}}{1+e^{A+Bx}}$

Set $\boldsymbol{A = 1}$

$\boldsymbol{B = -2}$



Logit / Logistic function

More importantly, in the following equation, we isolate p on one side.

$$\ln\left(\frac{p}{1-p}\right) = \underbrace{A + B * x}_{z}$$

where x is any predictor

Logit / Logistic function

$$logit(\boldsymbol{p}) = \ln\left(\frac{\boldsymbol{p}}{1-\boldsymbol{p}}\right)$$

$$\boldsymbol{p} = \boldsymbol{logistic}(\,A + Bx\,) = \frac{e^{\,A+B\boldsymbol{x}}}{1 + e^{A+B\boldsymbol{x}}}$$

In a standard regression model, we model **Y** as a **linear function** of the predictor **X**

$$Y \;=\; A + B\boldsymbol{x}$$

Here, we perform an additional mapping, where the **linear function** of **X** (A + B**x**)

is transformed into a **probability** using the logistic function.

That's why it is called "**Logistic Regression Model**"

Logit / Logistic function

$$\ln\left(\frac{p}{1-p}\right) = A + Bx$$

where x is any predictor (e.g. hours)

$$p = logistic(A + Bx)$$

$$= \frac{e^{A+Bx}}{1+e^{A+Bx}}$$

This expression is very helpful for estimating the **probability** based on the predictor (number of hours studied)

Logit / Logistic function

$$\ln\left(\frac{p}{1-p}\right) = A + Bx$$

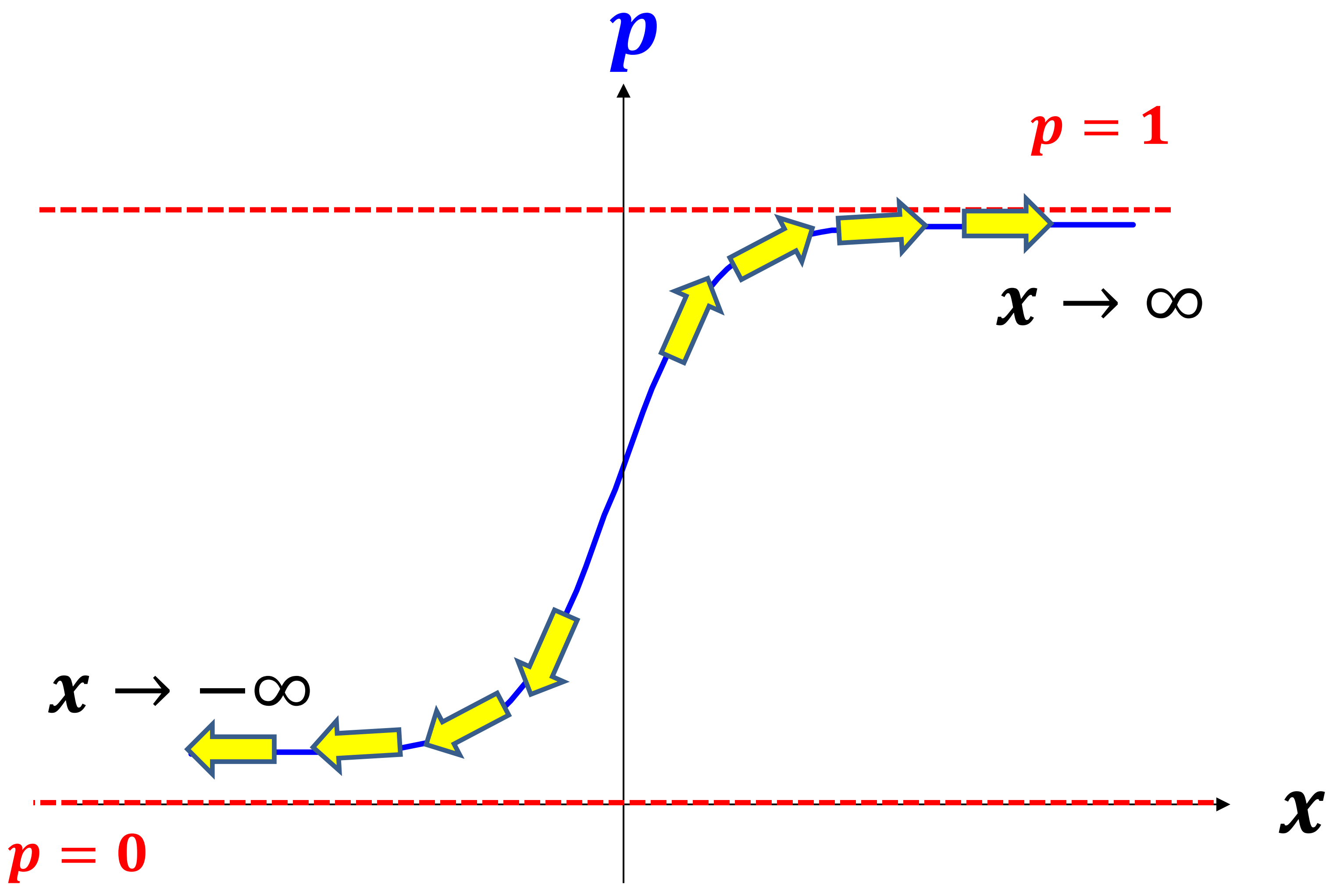
where x is any predictor (e.g. hours)

$$p = logistic(A + Bx)$$

$$= \frac{e^{A+Bx}}{1 + e^{A+Bx}}$$

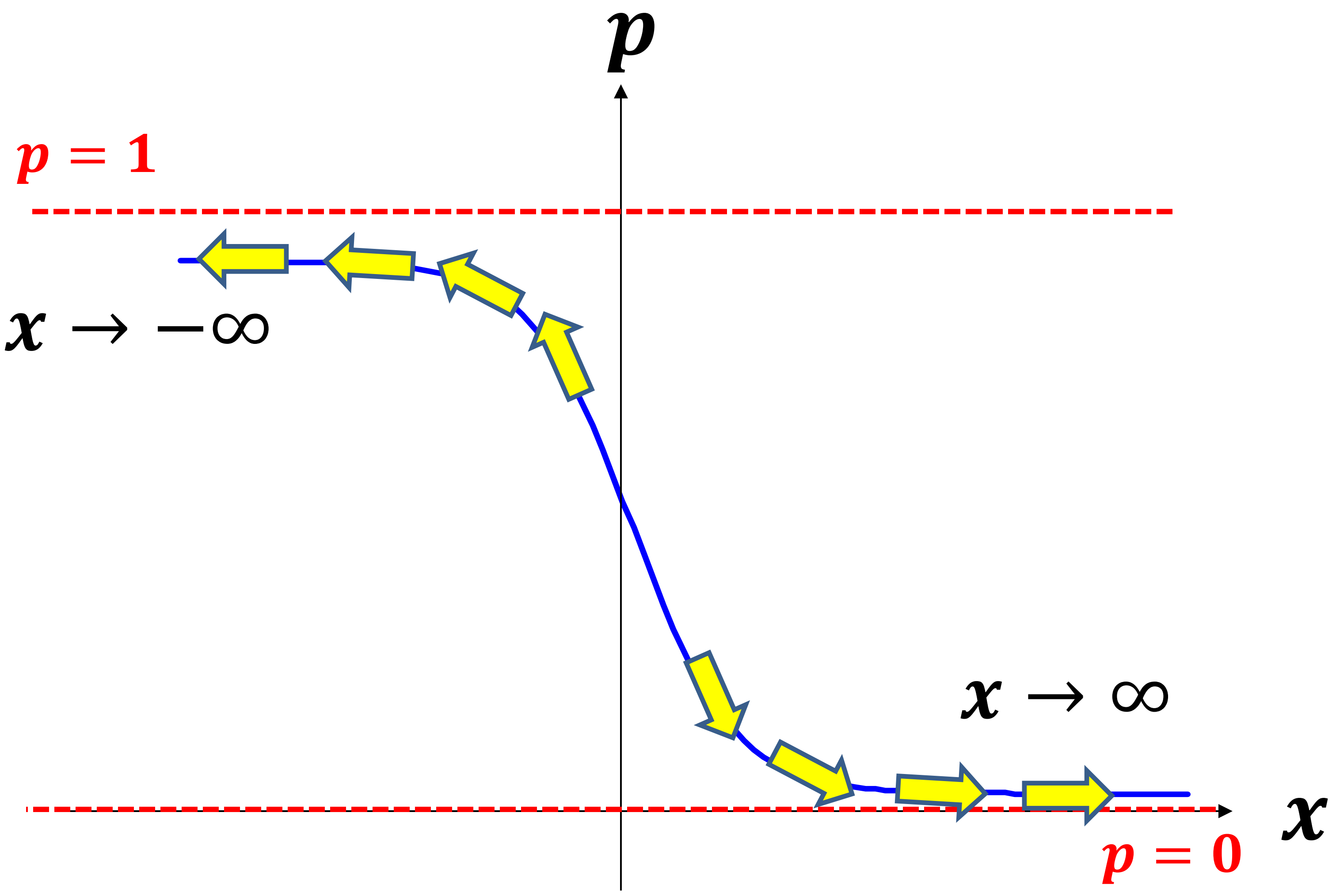
This expression is very helpful for estimating the **probability** based on the predictor (**number of hours studied**)

Let's graph the function $p = \dfrac{e^{A+Bx}}{1+e^{A+Bx}}$ Set $A = 1$, $B = 2$



Let's graph the function $p = \dfrac{e^{A+Bx}}{1+e^{A+Bx}}$ Set $A = 1$

$B = -2$



Additional Properties of the Logistic Regression Model

$$\ln\left(\frac{p}{1-p}\right) = A + B * x$$

where x is any predictor

Odds of an event (e.g. passing)