Logistic Regression: Formulation

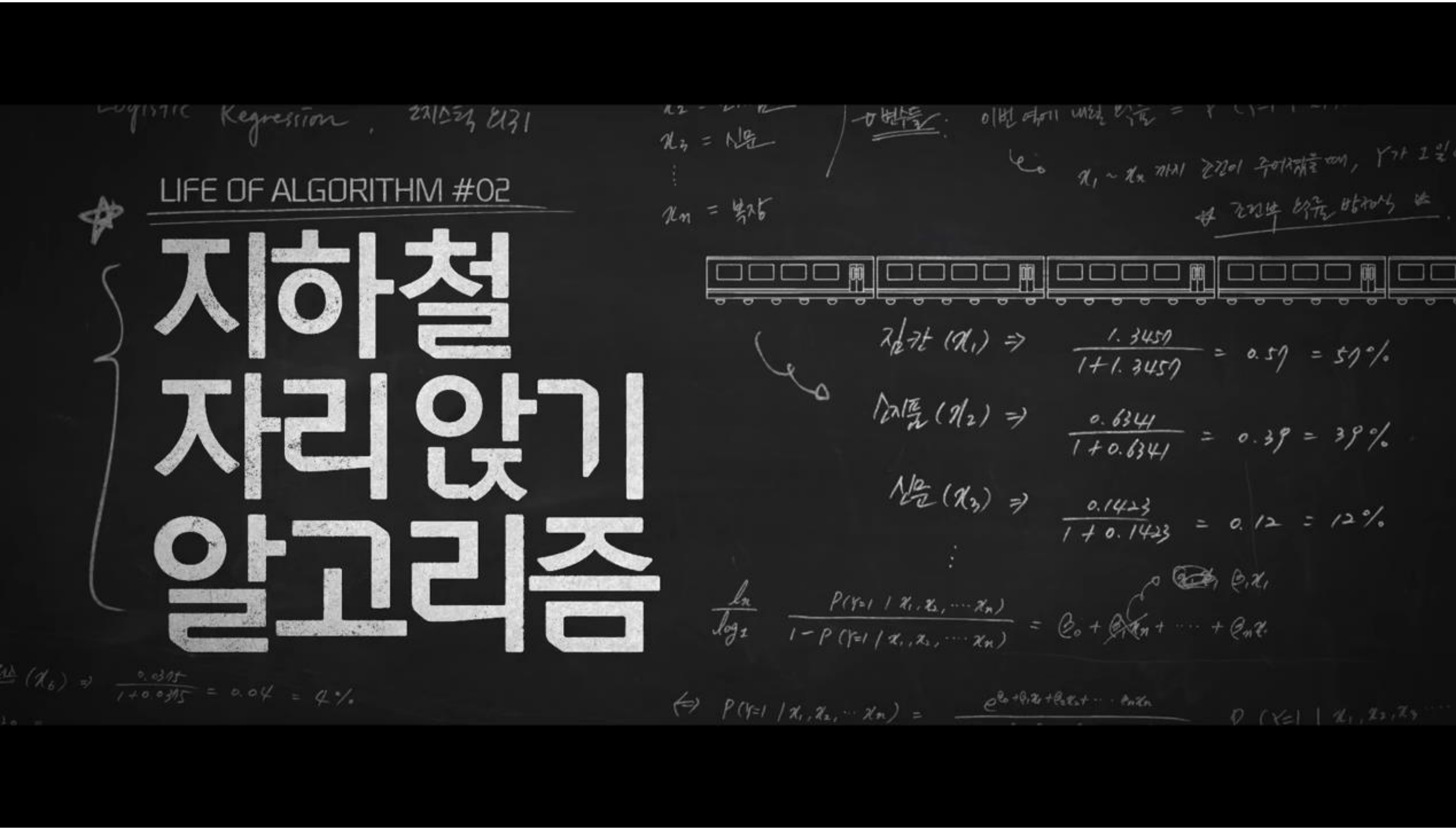
강필성

고려대학교 산업경영공학부

pilsung_kang@korea.ac.kr

AGENDA

Logistic Regression



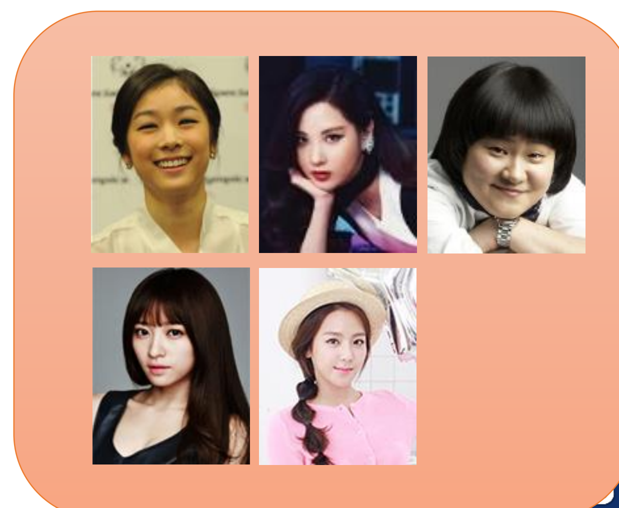
Logistic Regression

- Classification

Men

Vs.

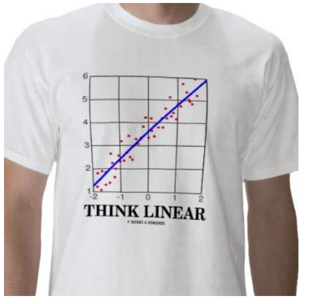
Women



Revisit Multiple Linear Regression

- Goal
  - ✓ Fit a linear relationship between a quantitative dependent variable Y and a set of predictors $X_1, X_2, \ldots, X_d$.

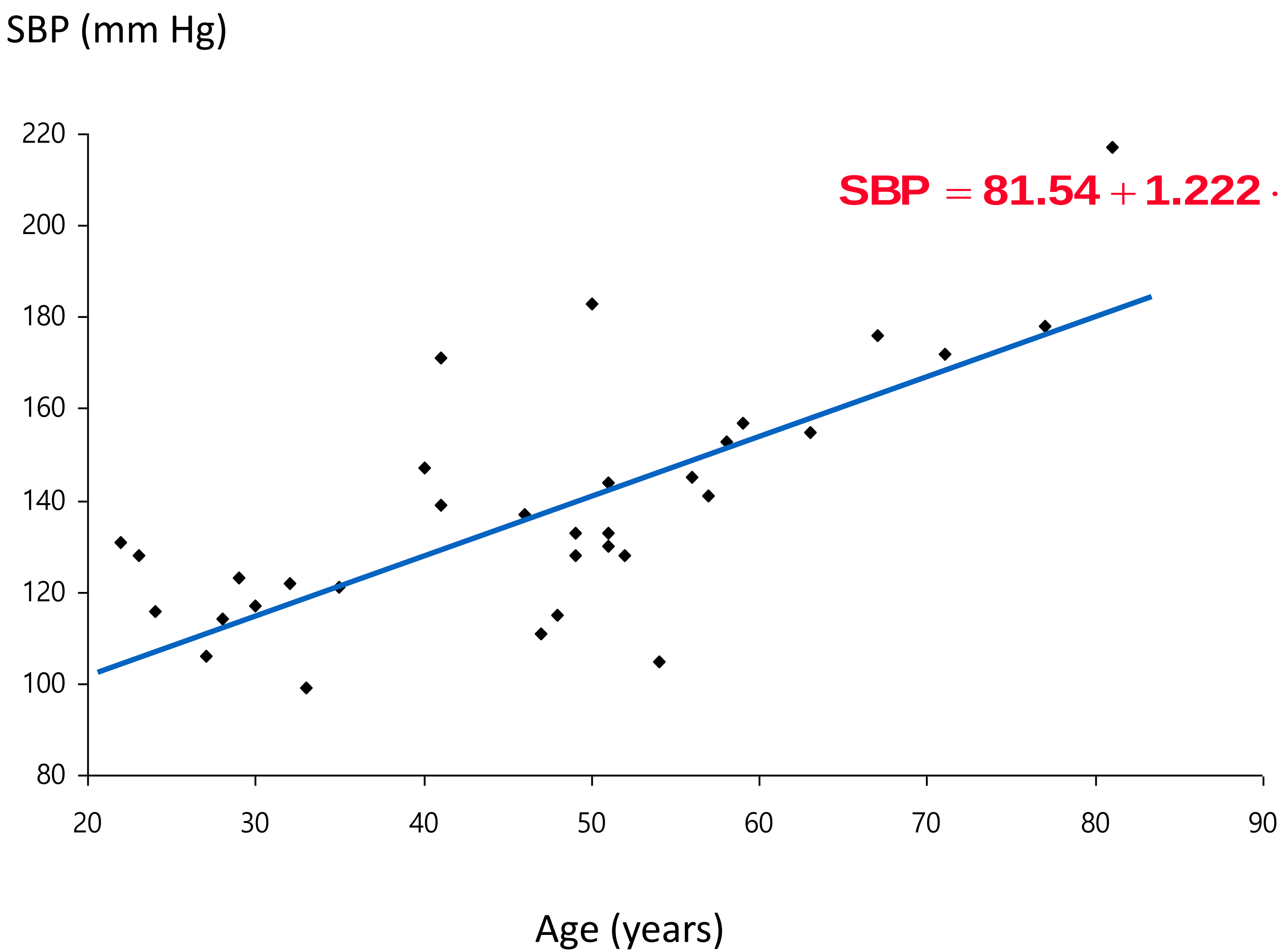
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d$$



- Example 1
  - ✓ Age and systolic blood pressure (SBP) among 33 adult women.

Age	SBP	Age	SBP	Age	SBP
22	131	41	139	52	128
23	128	41	171	54	105
24	116	46	137	56	145
27	106	47	111	57	141
28	114	48	115	58	153
29	123	49	133	59	157
30	117	49	128	63	155
32	122	50	183	67	176
33	99	51	130	71	172
35	121	51	133	77	178
40	147	51	144	81	217

Revisit Multiple Linear Regression



$$SBP = 81.54 + 1.222 \cdot Age$$

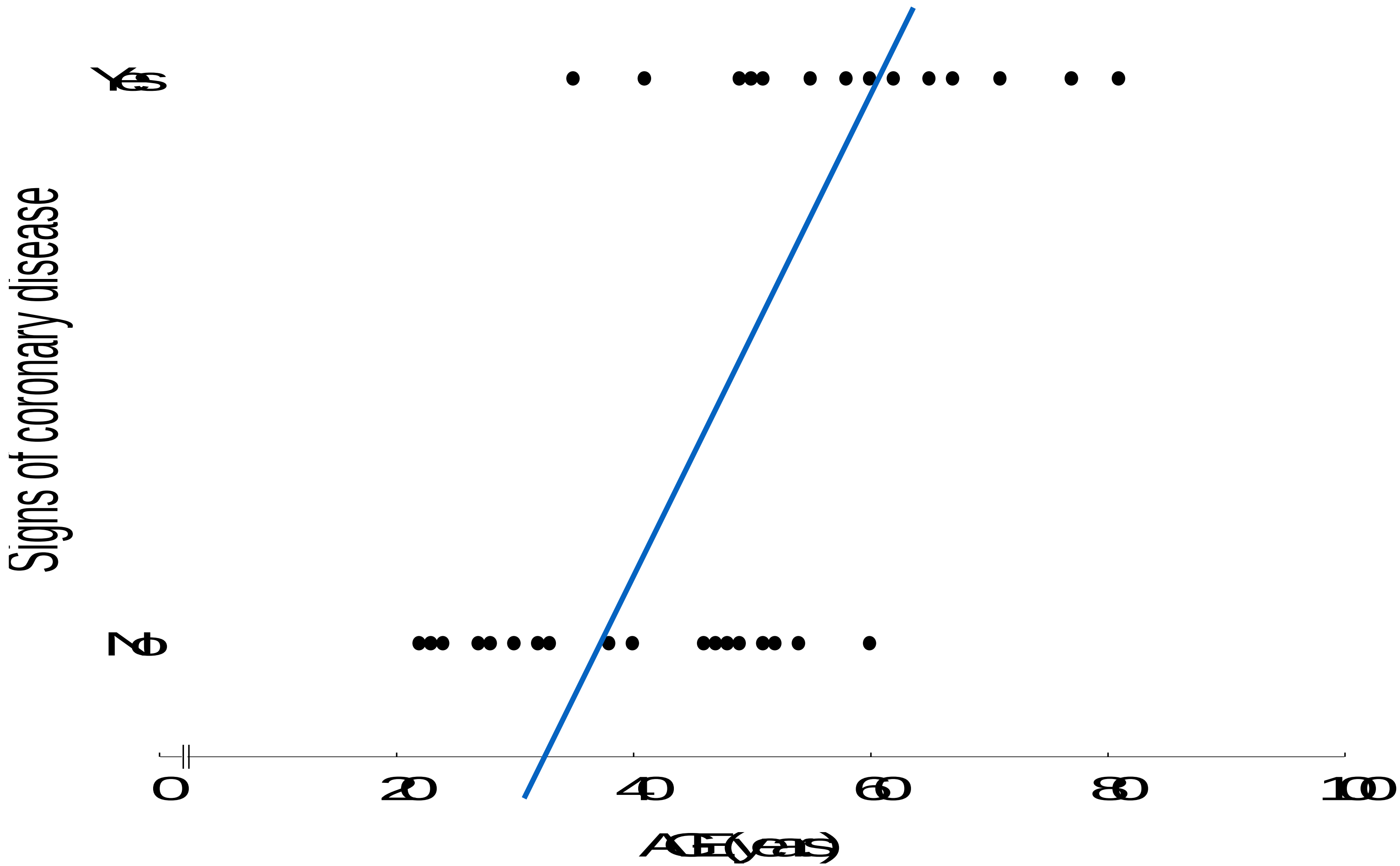
What If

- Example 2
  - ✓ Age and signs of coronary heart disease (CD)

Age	CD	Age	CD	Age	CD
22	0	40	0	54	0
23	0	41	1	55	1
24	0	46	0	58	1
27	0	47	0	60	1
28	0	48	0	60	0
30	0	49	1	62	1
30	0	49	0	65	1
32	0	50	1	67	1
33	0	51	0	71	1
35	1	51	1	77	1
38	0	52	0	81	1

What If

- Linear regression does not estimate Pr(Y=1|X) well

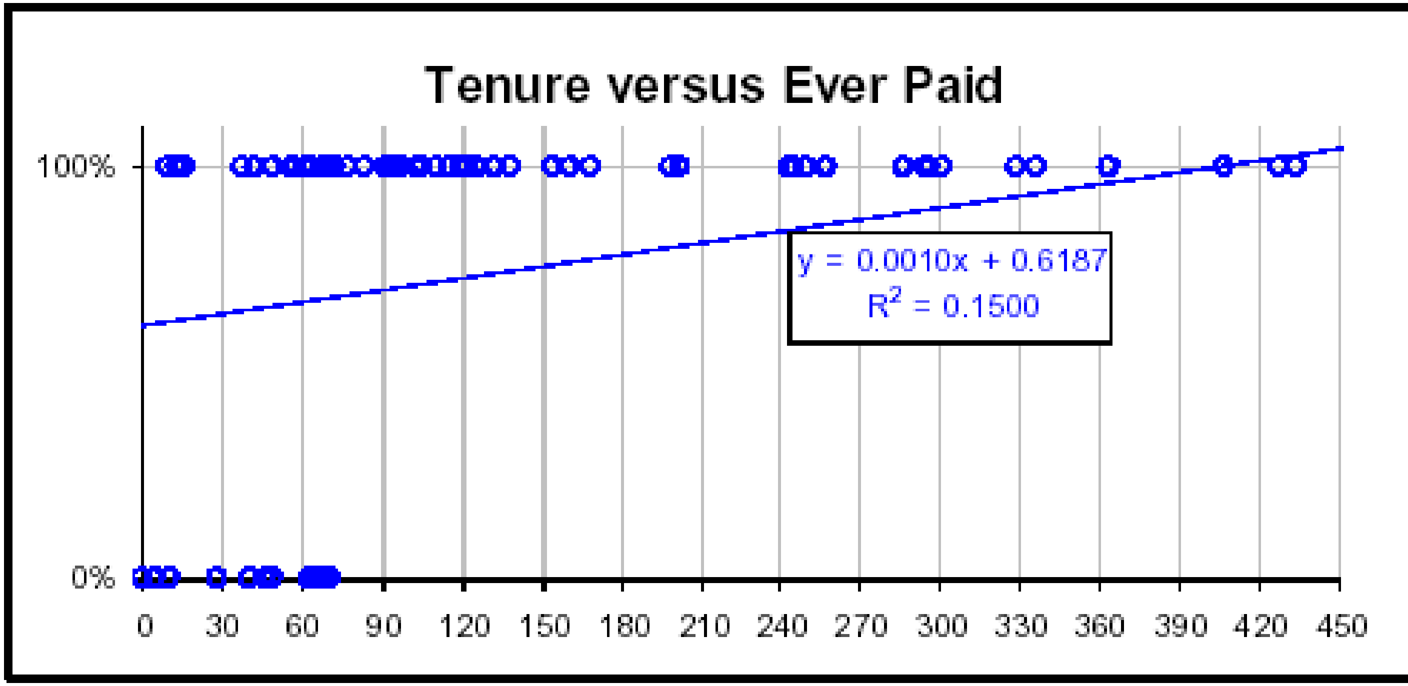


For Classification Task

- Is it appropriate to model the probability as a function of predictors?

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d$$

✓ May have a probability that is greater than 1 or less than 0



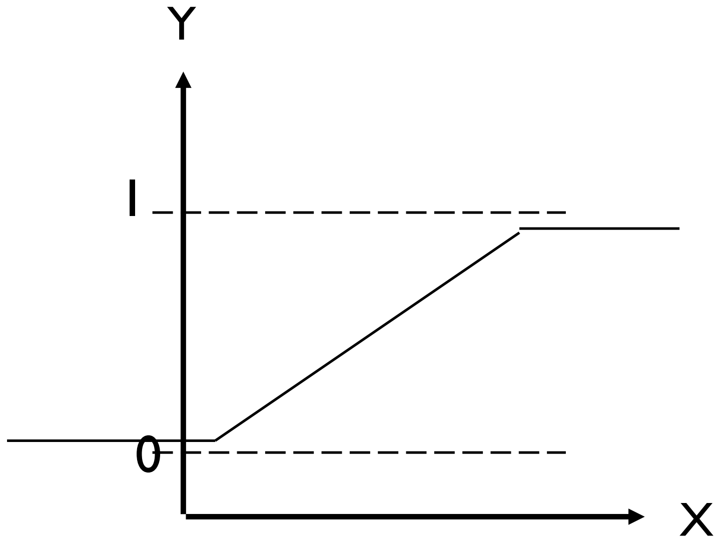
고려대학교 KOREA UNIVERSITY

DSBA Data Science & Business Analytics

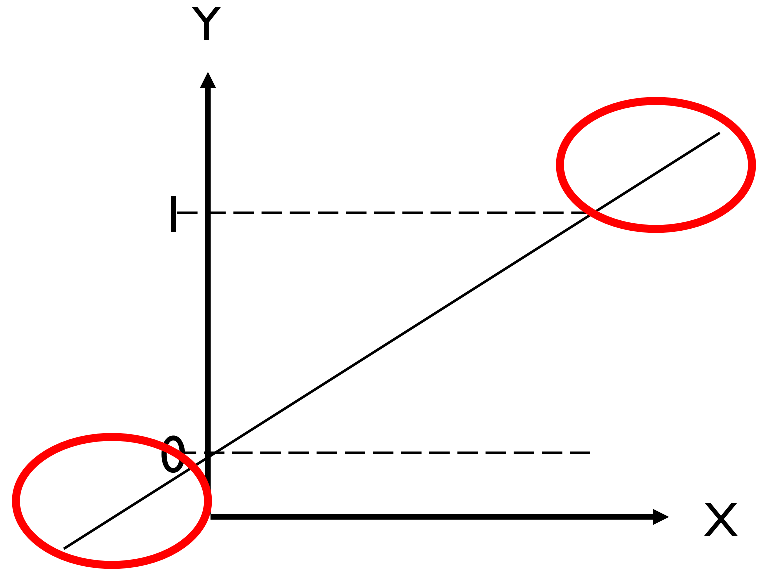
For Classification Task

- Consider when there are only two outcomes (0 & 1)
  - ✓ Is a linear model appropriate?

Ideally:

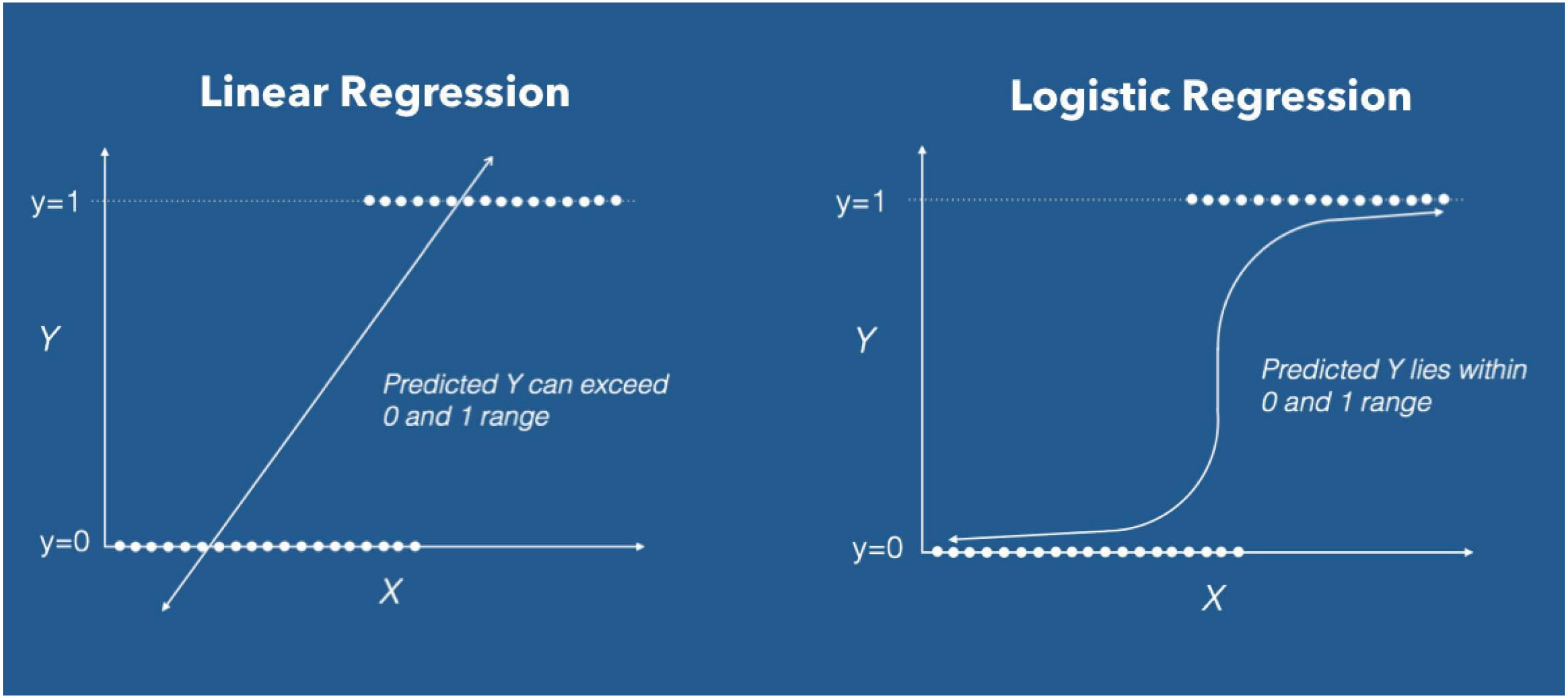


Reality:



For Classification Task

- Consider when there are only two outcomes (0 & 1)
  - ✓ Is a linear model appropriate?



https://medium.com/greyatom/logistic-regression-89e496433063

For Classification Task

- Problem
  - ✓ For binary classification tasks, there only two possible outcomes (0 and 1)
  - ✓ Regression equation has no limit on the generated value
  - ✓ Allowed ranges of the input X and the output y do not match

$$\hat{y} = \boxed{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d}$$

Only 0 or 1 are allowed

All real values are possible

  - ✓ Goal: Build a classification model that inherit the advantages of regression model (ability to find significant variables, explainability, etc)

고려대학교 KOREA UNIVERSITY

DSBA Data Science & Business Analytics

Logistic Regression: Goal

- Goal:
  - ✓ Find a function of the predictor variables that relates them to a 0/1 outcome

- Features:
  - ✓ Instead of Y as outcome variable (like in linear regression), we use a function of Y called the "logit".
  - ✓ Logit can be modeled as a linear function of the predictors.
  - ✓ The logit can be mapped back to a probability, which, in turn, can be mapped to a class.

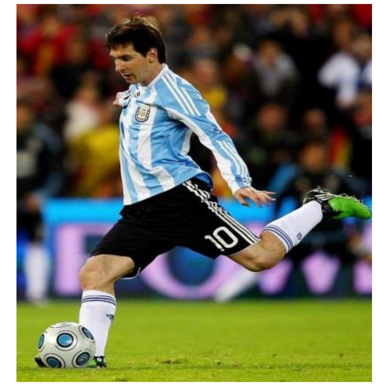
Logistic Regression: Odds

- 2010 World Cup Betting Odds


9 : 2


9 : 2


6 : 1


9 : 1


200 : 1


250 : 1


500 : 1


1000 : 1

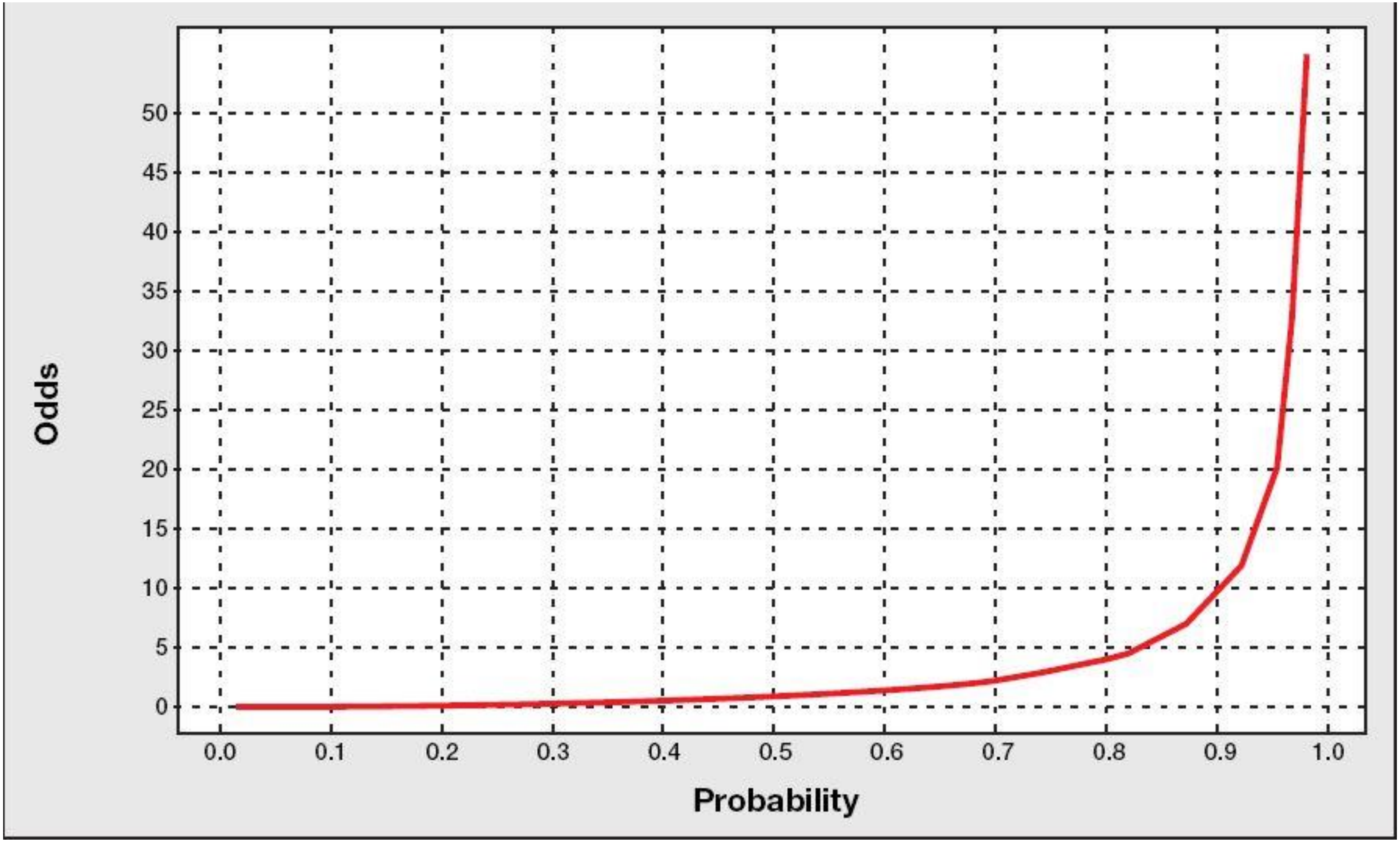
Logistic Regression: Odds

- Odds
  - ✓ p = probability of belonging to class 1 (success).

$$Odds = \frac{p}{1-p}$$

- For the previous examples
  - ✓ Winning odds of the Spain = 2/9, then the winning probability of the Spain = 2/11.
  - ✓ Winning odds of the Korea = 1/250, then the winning probability of the Korea = 1/251 ≒ 0.00398 (0.398%)

Logistic Regression: Odds



고려대학교 KOREA UNIVERSITY

DSBA Data Science & Business Analytics

Logistic Regression: Log odds

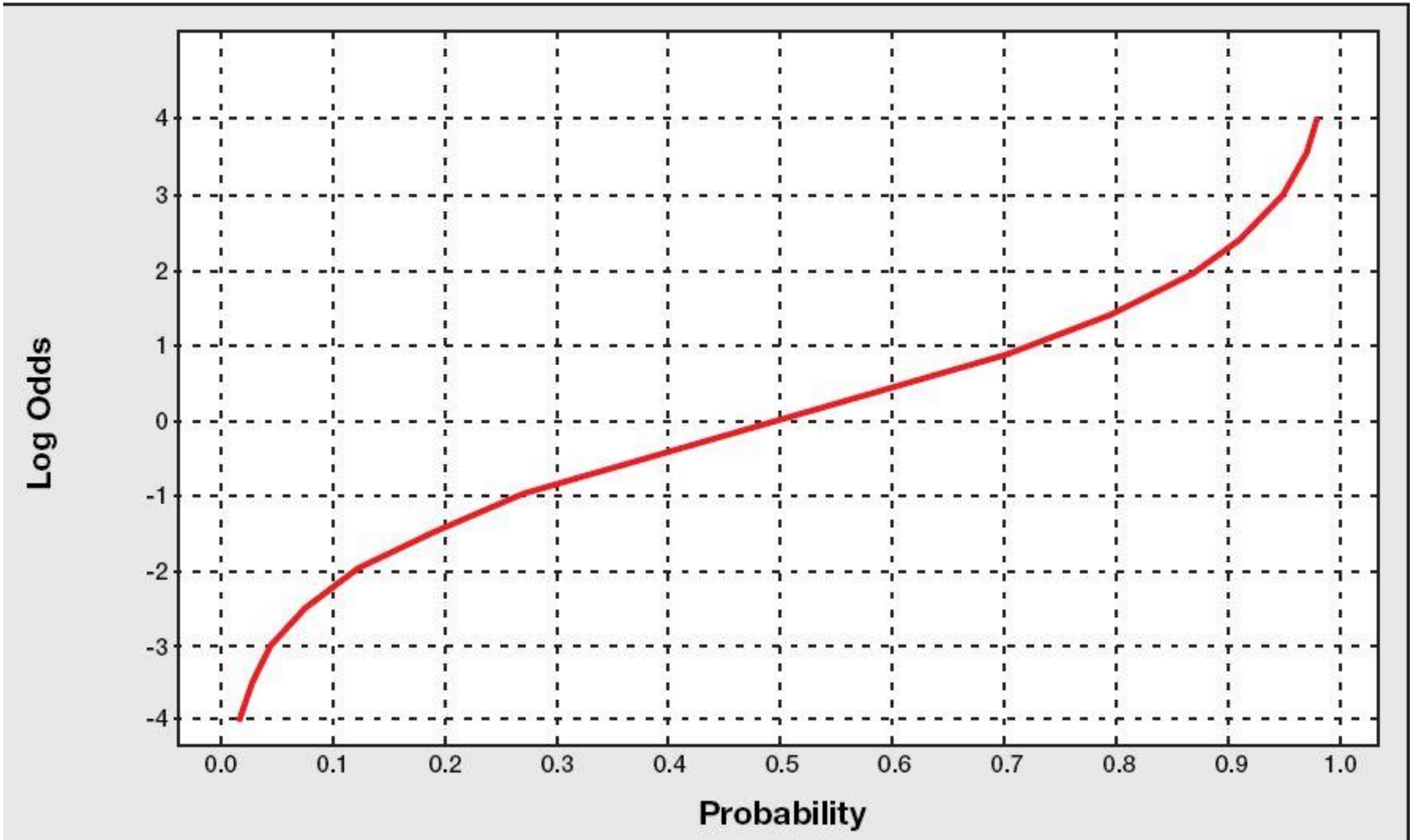
- The limitation of the odds
  - ✓ $0 < \text{odds} < \infty$
  - ✓ Asymmetric

- Take the logarithm of the odds

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

  - ✓ $-\infty < \log(\text{odds}) < \infty$
  - ✓ Symmetric
  - ✓ Negative when p is small and positive when p is large

Logistic Regression: Log odds



Logistic Regression: Equation

- Logistic regression equation
  - ✓ Linear equation for the odds:

$$log(Odds) = log\left(\frac{p}{1-p}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d$$

  - ✓ Take the exponential for the both sides:

$$\frac{p}{1-p} = e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d}$$

  - ✓ For the probability of the success:

$$p = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d)}} = \boxed{\sigma(\mathbf{x}|\beta)}$$

Logistic Regression: Equation

- Logistic regression equation

Logistic Regression 선형식

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

: logit
(odds에 자연로그를 취한 상태)

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i$$

- 로지스틱 회귀 모형은 종속변수가 이분형일 때 선형회귀모형의 제약을 극복하기 위해 확률에 대한 로짓 변환을 고려하여 분석

$$P = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i)}$$

- 위의 모형식에서 추정된 회귀계수로부터 사후확률에 대한 추정식을 계산


ANY
questions?