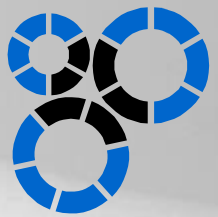
Machine Learning
CSE - 465

Lecture - 03

Outline

- What is Logistic Regression
- Use of Logistic Regression
- Logistic Curve
- Types of Logistic Regression
- The Logistic Regression Model
- The Odds Ratio
- Maximum Likelihood
- Linear Regression vs Logistic Regression



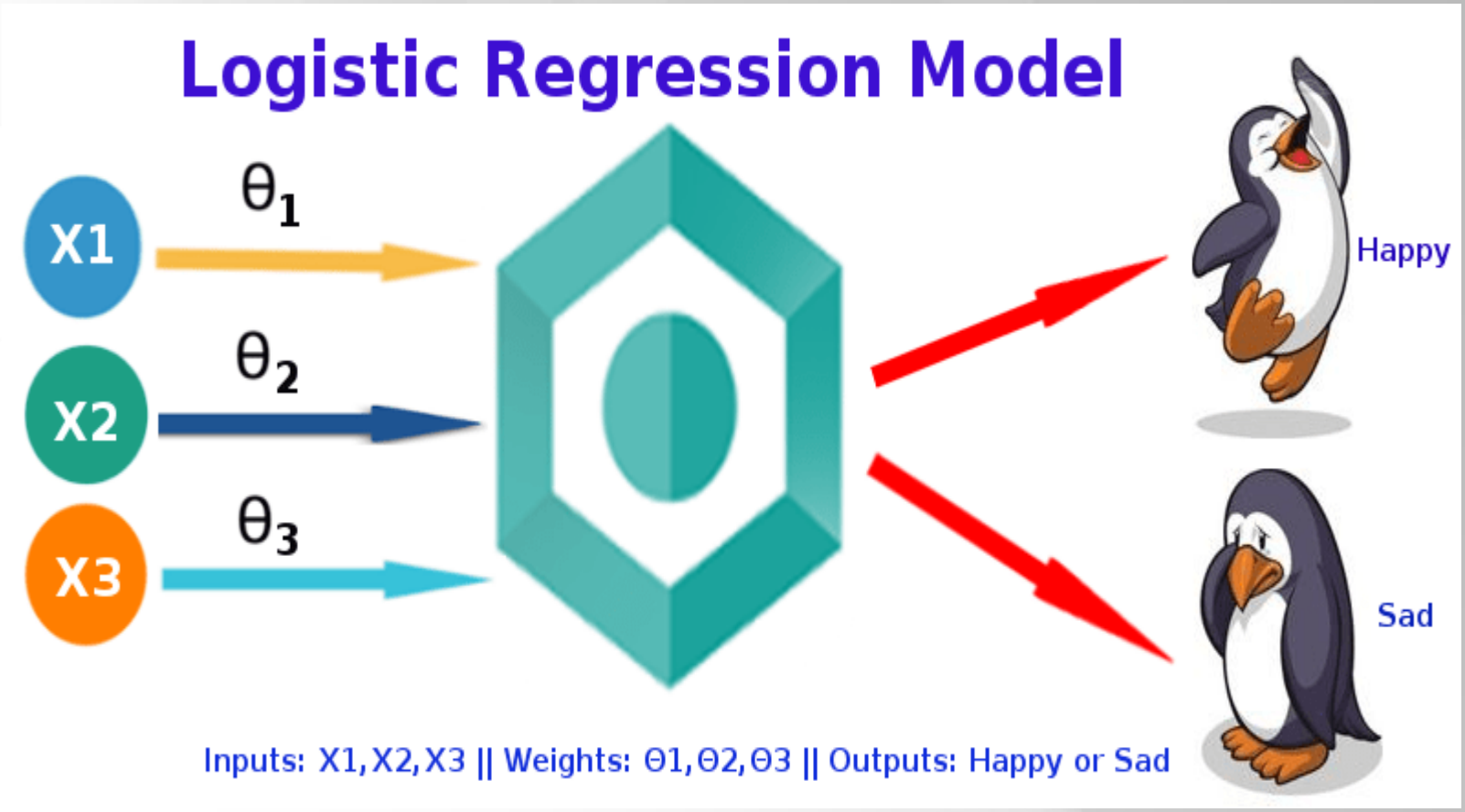
What is Logistic Regression

- Logistic regression is a form of regression analysis in which the outcome variable is binary or dichotomous

- Used when the research method is focused on whether or not an event occurred, rather than when it occurred (time course information is not used)

- Logistic Component
  - Instead of modeling the outcome, Y, directly, the method models the log odds(Y) using the logistic function.

What is Logistic Regression



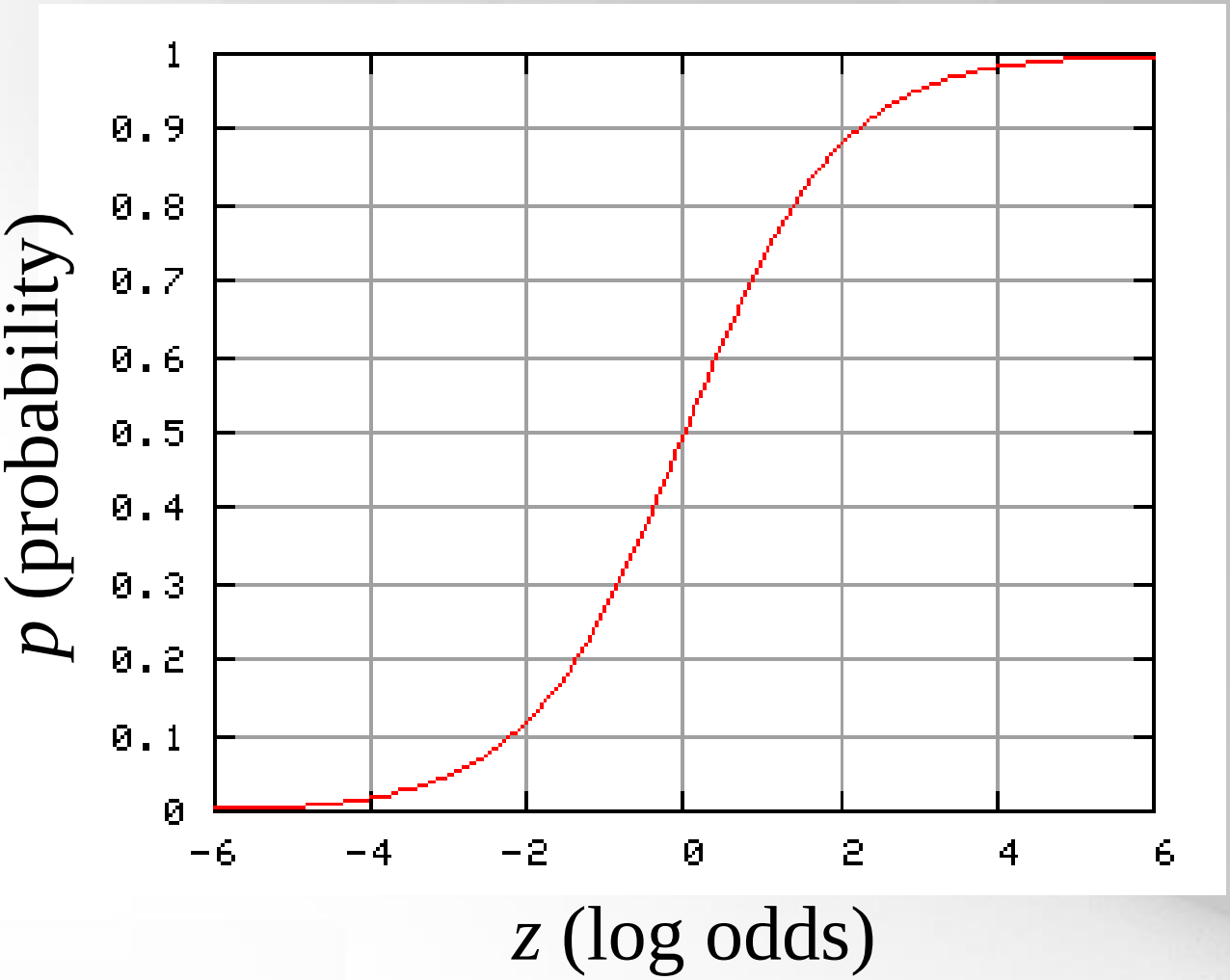
What can we use Logistic Regression for?

- To estimate adjusted prevalence rates, adjusted for potential confounders (socio-demographic or clinical characteristics)
- To estimate the effect of a treatment on a dichotomous outcome, adjusted for other covariates
- Explore how well characteristics predict a categorical outcome

Prevalence is a statistical concept referring to the number of cases of a disease that are present in a particular population at a given time, whereas **incidence** refers to the number of new cases that develop in a given period of time.

The Logistic Curve

$$LOGIT(p) = \ln\left(\frac{p}{(1-p)}\right) = z \quad \Leftrightarrow \quad p = \frac{\exp(z)}{1+\exp(z)}$$



Types of Logistic Regression

- Simple logistic regression
  - Logistic regression with 1 predictor variable
- Multiple logistic regression
  - Logistic regression with multiple predictor variables
  - Also known as multivariable logistic regression or multivariate logistic regression

The Logistic Regression Model

Logistic Regression:

$$\ln\left(\frac{\mathrm{P(Y)}}{\mathrm{1\text{-}P(Y)}}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_\mathrm{K} X_\mathrm{K}$$

Linear Regression:

$$\mathrm{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_\mathrm{K} X_\mathrm{K} + \varepsilon$$

The Logistic Regression Model

predictor variables

$$\ln\left(\frac{P(Y)}{1-P(Y)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K$$

dichotomous outcome

$\ln\left(\frac{P(Y)}{1-P(Y)}\right)$ is the log(odds) of the outcome.

The Logistic Regression Model

$$\ln\left(\frac{P(Y)}{1\text{-}P(Y)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K$$

intercept **model coefficients**

$$\ln\left(\frac{P(Y)}{1-P(Y)}\right)$$ is the log(odds) of the outcome.

Relationship between Odds & Probability

$$\text{Odds}(\text{event}) = \frac{\text{Probability}(\text{event})}{1\text{-Probability}(\text{event})}$$

$$\text{Probability}(\text{event}) = \frac{\text{Odds}(\text{event})}{1+\text{Odds}(\text{event})}$$

The Odds Ratio

- Odds Ratio is the ratio of two odd estimates.

P(response | male) = 0.40

P(response | female) = 0.20

$$\text{Odds (response | male)} = \frac{0.40}{1-0.40} = 0.667$$

$$\text{Odds (response | female)} = \frac{0.20}{1-0.20} = 0.25$$

$$\text{Odds Ratio (male: female)} = \frac{0.667}{0.25} = 2.67$$

The Odds Ratio

- An Odds Ratio of 2.67 for male vs female does not mean that the outcome is 2.67 times as likely to occur.

- It means that the odds of the outcome occurring are 2.67 times as high for male vs female

Assumptions in Logistic Regression

- Y_i are from Bernoulli or binomial (n_i, μ_i) distribution
- Y_i are independent
- Log odds $P(Y_i = 1)$ or logit $P(Y_i = 1)$ is a linear function of covariates

A **covariate** is a variable that is possibly predictive of the outcome under study

Maximum Likelihood

Idea of Maximum Likelihood

- Flipped a fair coin 10 times:

  T, H, H, T, T, H, H, T, H, H

- What is the *P(Heads)* given the data?

  1/100? 1/5? 1/2? 6/10?

$$\hat{p} = \frac{X}{N} = \frac{\#\text{ of heads}}{\text{total}\,\#\text{ of flips}}$$

Maximum Likelihood

- Maximum Likelihood Estimation (MLE) is a method of estimating the parameters of a probability distribution by maximizing a likelihood function, so that under the assumed statistical model the observed data is most probable.

- Standard errors are obtained as a by-product of the maximization process

Why not use linear regression for dichotomous outcomes?

- If we model Y directly and Y is dichotomous, this necessarily violates the linear regression assumption

- One of the more intuitive reasons not to is that will end up with predicted Y's other than 0 or 1 (possibly more extreme than 0 or 1).

Linear Regression vs Logistic Regression

Linear Regression	Logistic Regression
A linear approach that models the relationship between a dependent variable and one or more independent variables.	A statistical model that predicts the probability of an outcome that can only have two values.
Used to solve regression problems	Used to solve classification problems (binary classification).
Estimates the dependent variable when there is a change in the independent variable.	Calculates the possibility of an event occurring.
Output value is continuous.	Output value is discrete.
Uses a straight line.	Uses an S curve or sigmoid function.
Example: predicting the GDP of a country, predicting product price, predicting the house selling price, score prediction.	Example: Predicting whether an email is spam or not, whether a credit card transaction is fraud or not, whether a customer will take a loan or not.

Thank You