TOPIC 4. LOGISTIC REGRESSION

Introduction

- Logistic regression is a method that allows the prediction (classification) of discrete variables (outputs) based on a mixture of continuous and discrete predictors (inputs).

- It offers a non-linear relationship between the outputs and inputs (but linear after certain transformation– generalized linear models)
  - *Example: the probability of heart disease changes very little with a ten-point difference among people with low-blood pressure, but a ten point change can mean a drastic change in the probability of heart disease in people with high blood-pressure.*

Introduction

- Application criteria: the limitation on using the logistic regression is that the output should be discrete.

Terms used in logistic regression

- Odds - like probability, a measure of the likelihood of an event.
  - *Odds are usually written as “1 to 4 odds” which is equivalent to 1 out of five or .20 probability or 20% chance.*

- Odds ratio – the ratio of the odds over 1 – the odds. The probability of winning over the probability of losing. 1 to 4 odds equates to an odds ratio of .20/.80 = .25.

- Logit – this is the natural log of an odds ratio; often called a log odds.

Introduction

- *Y* = A BINARY RESPONSE (Discrete Variable), p and $1 - p$
  - *1 POSITIVE RESPONSE (Success)* → p
  - *0 NEGATIVE RESPONSE (failure)* → $1 - p$
- MEAN(*Y*) = p, observed proportion of successes
- VAR(*Y*) = $p(1 - p)$, maximized when p = .50, variance depends on mean
- *X* = ANY TYPE OF PREDICTOR → Continuous, Dichotomous, Polytomous.
- p depends on *X*.

The Logistic Function

$$p(X) = \frac{e^u}{1 + e^u}$$

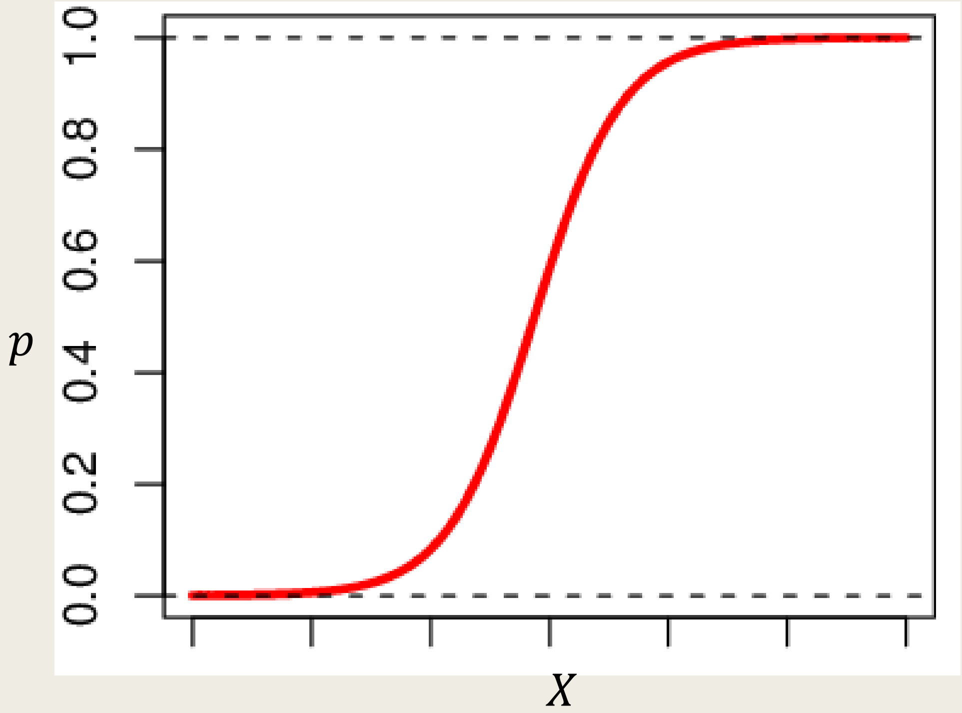
- Where $p(X)$ is the estimated probability that Y is in the success category and u is the regular linear regression equation:

$$u = b_0 + b_1 X$$

The Logistic Function

- Simple case

$$p(X) = \frac{e^{b_0 + b_1 X}}{1 + e^{b_0 + b_1 X}}$$



The Logit Function

- By algebraic manipulation, the logistic regression equation can be written in terms of an odds ratio for success:

$$\frac{p(X)}{1-p(X)} = e^{b_0+b_1X}$$

- Taking the natural log of both sides, we can write the equation in terms of logits (log-odds):

$$\ln\frac{p(X)}{1-p(X)} = b_0 + b_1X$$

Logistic Regression

- The response (binary)

$$Y \sim Bernoulli(p(X))$$

- Probability of success:

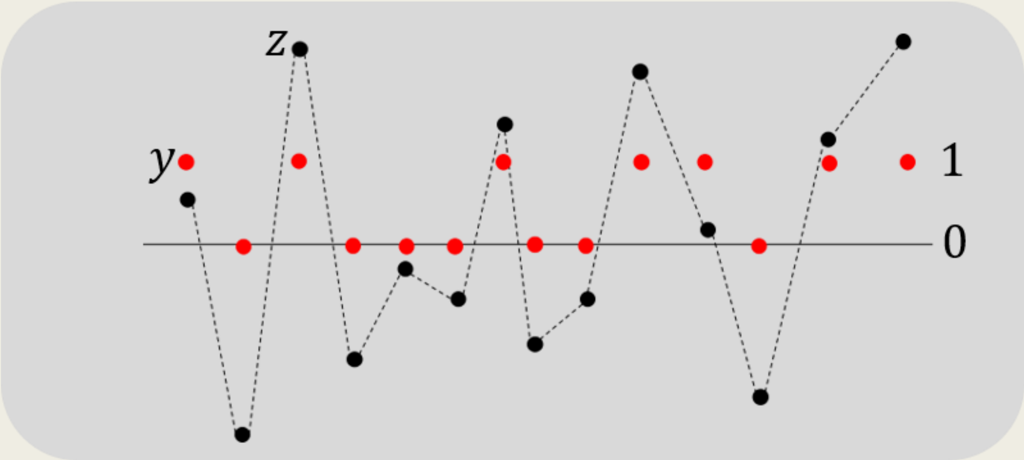
$$p(X) = \frac{e^{b_0 + b_1 X}}{1 + e^{b_0 + b_1 X}} \qquad \ln \frac{p(X)}{1 - p(X)} = b_0 + b_1 X$$

Latent Variable Model

$$Z = b_0 + b_1 X + \varepsilon$$

$$Y = \begin{cases} 1, & if\ Z > 0 \\ 0, & if\ Z \leq 0 \end{cases}$$

- Where the random error ε follows a logistic distribution.
- This latent variable model is equivalent to the standard logistic regression.



Interpretation

$$p(X) = \frac{e^{b_0+b_1X}}{1+e^{b_0+b_1X}} \qquad \ln\frac{p(X)}{1-p(X)} = b_0 + b_1X$$

- If $b_1 = 0$, there is no relationship between the response and the predictor.
- If $b_1 > 0$, when X gets larger so does the probability of success.
- If $b_1 < 0$, when X gets larger, the probability of success gets smaller.
- How much bigger or smaller depends on value of the slope.

Estimating Coefficients

- Given a dataset $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, find estimates for b_0, b_1.
- Maximum likelihood estimation (MLE): find $\hat{b}_0$, $\hat{b}_1$ that maximize the likelihood function

$$L(b_0, b_1) = \prod_{i=1}^{n} P(y_i) = \prod_{i: y_i = 1} P(y_i = 1) \prod_{i': y_{i'} = 0} P(y_{i'} = 0)$$

$$P(y_i = 1) = \frac{e^{b_0 + b_1 x_i}}{1 + e^{b_0 + b_1 x_i}}$$

$$P(y_{i'} = 0) = \frac{1}{1 + e^{b_0 + b_1 x_{i'}}}$$

Example

- Assume the dataset has 5 samples

Patient	Risk score X	Surgery Outcome Y (survival/0;death/1)
1	3	0
2	10	1
3	6	1
4	8	1
5	0	0

$$P(y_1 = 0) = \frac{1}{1 + e^{b_0 + 3b_1}}$$

$$P(y_5 = 0) = \frac{1}{1 + e^{b_0}}$$

$$P(y_2 = 1) = \frac{e^{b_0 + 10b_1}}{1 + e^{b_0 + 10b_1}}$$

$$P(y_3 = 1) = \frac{e^{b_0 + 6b_1}}{1 + e^{b_0 + 6b_1}}$$

$$P(y_4 = 1) = \frac{e^{b_0 + 8b_1}}{1 + e^{b_0 + 8b_1}}$$

$$L(b_0, b_1) = \frac{1}{1 + e^{b_0 + 3b_1}} \cdot \frac{1}{1 + e^{b_0}} \cdot \frac{e^{b_0 + 10b_1}}{1 + e^{b_0 + 10b_1}} \cdot \frac{e^{b_0 + 6b_1}}{1 + e^{b_0 + 6b_1}} \cdot \frac{e^{b_0 + 8b_1}}{1 + e^{b_0 + 8b_1}}$$

Significance Test

- Hypothesis testing

H_0: $b_1 = 0$ (there is a relationship between *X* and *Y*)

H_1: $b_1 \neq 0$ (no relationship between *X* and *Y*)

- Z (Wald) test

$$W = \frac{\hat{b}_1}{SD(\hat{b}_1)} \sim N(0, 1)$$

Extension

- Multiple logistic regression

$$\ln \frac{p(X)}{1-p(X)} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k$$

- Multinomial logistic regression ($Y = 0, 1, 2, \ldots$)

$$\ln \frac{p_{Y=1}(X)}{p_{Y=0}(X)} = \cdots \quad \ln \frac{p_{Y=2}(X)}{p_{Y=0}(X)} = \cdots \quad \ln \frac{p_{Y=3}(X)}{p_{Y=0}(X)} = \cdots$$

- Ordinal (ordered) logistic regression ($Y = 0, 1, 2, \ldots$)

$$\ln \frac{p_{Y\le 0}(X)}{1-p_{Y\le 0}(X)} = \cdots \quad \ln \frac{p_{Y\le 1}(X)}{1-p_{Y\le 1}(X)} = \cdots \quad \ln \frac{p_{Y\le 2}(X)}{1-p_{Y\le 2}(X)} = \cdots$$

Generalized Linear Models

- Logistic regression is a special type of GLM.
- General form of GLMs:

$$Y \sim Probability\ Distribution(\mu)$$
$$g(\mu) = b_0 + b_1 X$$

Example: Logistic regression

$$Y \sim Bernoulli(p)$$
$$g(p) = ln\frac{p}{1-p} = b_0 + b_1 X$$

- g is called link function. It transforms the mean of the probability distribution to a linear function of X.
- $E(Y|X) = g^{-1}(b_0 + b_1 X)$ vs. $E(Y|X) = b_0 + b_1 X$ (linear regression)

Probit Regression

- Probit regression is another popular model for binary responses.
- Probit model:

$$Y \sim Bernoulli(p)$$

$$p = \Phi(b_0 + b_1 X)$$

$$g(p) = \Phi^{-1}(p) = b_0 + b_1 X$$

- Φ is the cumulative distribution of standard normal distribution.

Latent Variable Model of Probit Regression

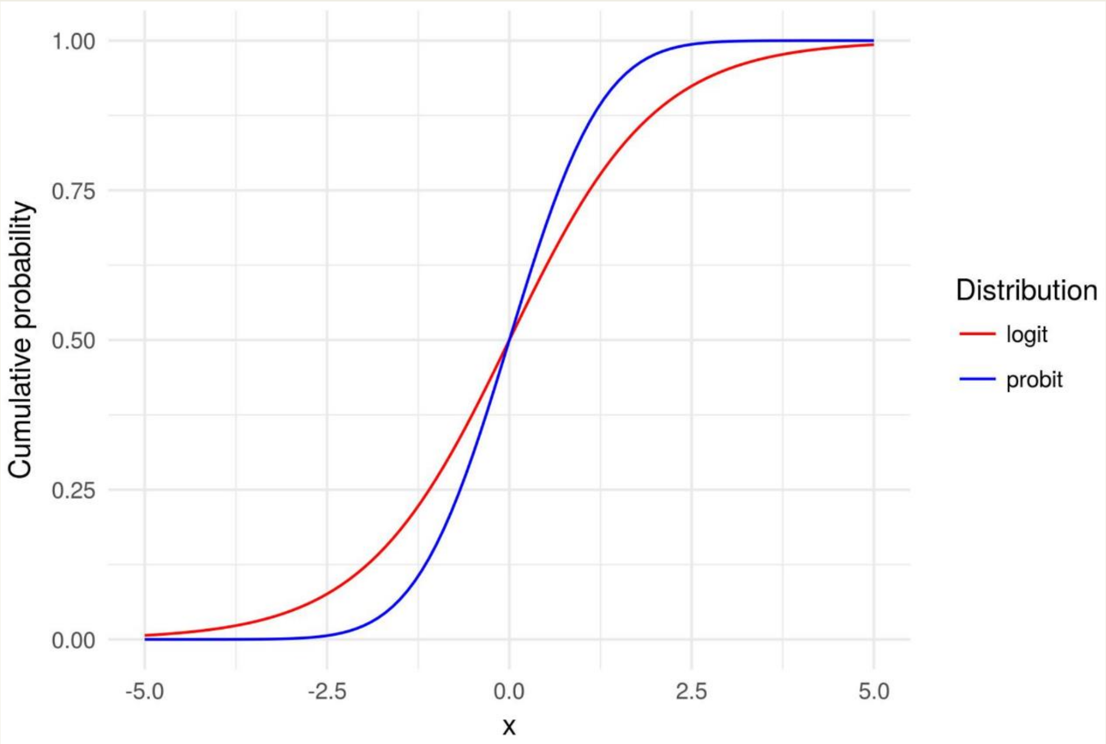
$$Z = b_0 + b_1 X + \varepsilon$$

$$Y = \begin{cases} 1, & if\ Z > 0 \\ 0, & if\ Z \le 0 \end{cases}$$

- Where the random error ε follows a standard normal distribution. So the model of Z is a linear regression model.
- This latent variable model is equivalent to the probit regression.

Logit vs. Probit Regression

- Logit function and probit function have similar shapes.



(from datacamp.com)

Summary

- Generalized linear models including logistic regression is an extension of probability models to incorporate the effects of predictors.
- These models are easy to understand and have good interpretation.