Logistic regression

- In may cases, we may have a binary depedent variable: That is,

$$Y_i = \begin{cases} 1 \\ 0 \end{cases} \text{ , e.g., approval vs denial, employed vs unemployed, pass vs fail...}$$

- Mortgage application : outcome is either approval or denial. This decision is made based on several factors such as income, job, debt, ... Suppose that we want to examine the relation between this mortgage approval decision and loan payment/income (P/I ratio). Let

$$Y_i = \begin{cases} 1, \text{ if denied} \\ 0, \text{ if approved} \end{cases}, \; X_i = \text{P/I ratio},$$

- We are interested in the probability of approval given my P/I ratio, $P\left(\text{denial} \mid \text{P/I ratio}\right)$.

- Suppose that this decision is based on latent variable Y^* (e.g., risk score) which is determined by X_i =P/I ratio and other unobserved factors, such that

$$Y^* = \beta_0 + \beta_1 X_i + \varepsilon_i$$

and

$$Y_i = \begin{cases} 1, \text{ if } Y^* \geq 0 \\ 0, \text{ if } Y^* < 0 \end{cases}.$$

- Here we assume that $\varepsilon_i | X_i$ follows the logistic distribution.

$$L\left(\varepsilon | X\right) = \frac{1}{1+\exp\left(-\varepsilon\right)},$$

where $L\left(\cdot | X\right)$ is the conditional cdf of the logistic random variable.

- Note that

$$\begin{aligned} P\left(Y_i = 1 | X_i\right) &= P\left(Y_i^* \geq 0 | X_i\right) \\ &= P\left(\beta_0 + \beta_1 X_i + \varepsilon_i \geq 0 | X_i\right) \\ &= P\left(\varepsilon_i \geq -\beta_0 - \beta_1 X_i | X_i\right) \\ &= P\left(\varepsilon_i \leq \beta_0 + \beta_1 X_i | X_i\right) \\ &= L\left(\beta_0 + \beta_1 X_i | X_i\right) \\ &= \frac{1}{1+\exp\left(-\beta_0 - \beta_1 X_i\right)} \end{aligned}$$

- We can estimate the coefficients using "maximum likelihood estimation" (MLE), and the estimated probability of approval given the P/I ratio is

$$\hat{P}\left(Y_i = 1 | X_i\right) = \frac{1}{1+\exp\left(-\hat{\beta}_0 - \hat{\beta}_1 X_i\right)}.$$

- Example: Suppose that $\hat{\beta}_0 = -2, \hat{\beta}_1 = 3, X = 0.4$,

$$P\left(Y = 1 | X = 0.4\right) = \left(= -2 + 3 \times 0.4\right) = L\left(-0.8\right).$$

- Properties

  1. It has S-hsape, $0 \leq P\left(Y = 1 | X\right) \leq 1$ and $P\left(Y = 1 | X\right)$ increases in X when $\beta_1 > 0$.

  2. β_1 does not represent the change in $P\left(Y = 1 | X\right)$ by increasing 1 unit increase in X. Note that the cdf has an S shape. The marginal probability is the function of X.

- Predicted probabilty
$$\hat{P}\left(Y=1|X\right)=L\left(\hat{\beta}_0+\hat{\beta}_1X\right).$$
- Estimation: OLS does not work! We cannot use the lm() function!