1 Logistic Regression

1.1 The Logistic Regression Model

- When the response variable Y is **dichotomous** (i.e. only takes values **zero or one**), ordinary linear regression models (that assume normality of the errors) aren't appropriate.

  Instead, we use the *logistic regression model.*

- Recall that a ***Bernoulli(π)*** random variable takes values **zero and one** with probabilities $\mathbf{1-\pi}$ and $\boldsymbol{\pi}$, respectively. Such random variables arise when individuals are classified into **two categories**, ***success*** and ***failure***, say, and we define

$$Y = \begin{cases} 1 & \text{if the individual is a } \textit{success} \\ 0 & \text{if the individual is a } \textit{failure} \end{cases} \tag{1}$$

  with

$$\begin{aligned} P(Y=1) &= \pi \\ P(Y=0) &= 1-\pi \end{aligned}$$

- Suppose $\boldsymbol{Y_1, Y_2, \ldots, Y_n}$ are independent **Bernoulli($\boldsymbol{\pi_i}$)** random variables. Note that $\boldsymbol{\pi_i = p(Y_i = 1)}$ is allowed to **differ** from **one individual** to the **next**.

  The mean response for the ith individual is

$$\mathrm{E}(Y_i) = 1 \cdot \pi_i + 0 \cdot (1-\pi_i) = \pi_i$$

  Thus the **mean response** is also the **probability** that the response will equal **one**.

  In *logistic regression*, we'll model the mean response as a function of a predictor variable X. It *won't* make sense to model the mean response as $\mathrm{E}(Y_i) = \beta_0 + \beta_1 X_i$, as was done for simple linear regression, because this wouldn't constrain $\mathrm{E}(Y_i) = \pi_i$ to lie between zero and one.

- The ***logistic regression model*** is

  Logistic Regression Model: Suppose $Y_1, Y_2, \ldots, Y_n$ are independent

Bernoulli(π_i) random variables, so

$$\mathrm{E}(Y_i) \;=\; P(Y_i = 1) \;=\; \pi_i \,.$$

The ***logistic regression model*** is

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) \;=\; \beta_0 + \beta_1 X_i \tag{2}$$

which (by exponentiating both sides and solving for π_i) can be written as

$$\pi_i \;=\; \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}} \tag{3}$$

where for $i = 1, 2, \ldots, n$,

- ▷ X_i is the value of the predictor variable X for the ith individual
- ▷ $\pi_i \;=\; P(Y_i = 1)$ is the probability that the ith individual's response is one

and β_0 and β_1 are parameters of the model.

- The function

$$g(\pi) \;=\; \log\left(\frac{\pi}{1-\pi}\right)$$

is an example of what's called a ***link function*** because it "links" the mean response $\mathrm{E}(Y) = \pi$ to the predictor X via the linear function $\beta_0 + \beta_1 X$ in (2).

More precisely, it's called the ***logit link***, and is the most widely use link function (for *logistic regression*). Another commonly used one is the ***probit link***. See the textbook.

- As a function of X, the **mean response function**

$$\pi(X) \;=\; \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

of the **logistic regression model** (3) is graphed in Fig. 1 for several values of β_0 and β_1. It has the following properties:

1. π is constrained to lie between zero and one.

2. If $\beta_1 > 0$, then π is an increasing function of X, and as $X \to \infty$, $\pi \to 1$, but as $X \to -\infty$, $\pi \to 0$.
3. If $\beta_1 < 0$, then π is decreasing function of X, and as $X \to \infty$, $\pi \to 0$, but as $X \to -\infty$, $\pi \to 1$.
4. The parameter β_1 determines how "steep" the "middle" part of the graph of π is as a function of X. A larger value of β_1 results in a "steeper" middle part of the graph.
5. The parameter β_0 determines the horizontal location of the "middle" part of the graph of π is as a function of X.

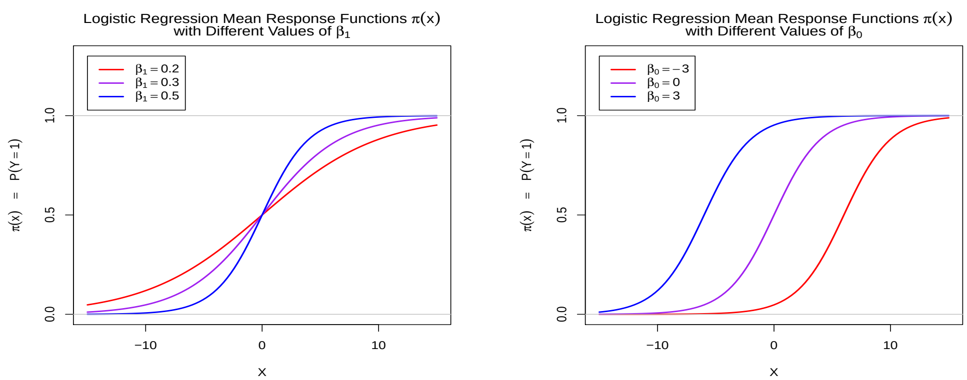

Figure 1

- The parameters β_0 and β_1 are estimated *not* by least squares, but by the ***maximum likelihood*** method.

- Once the **maximum likelihood** estimates $\boldsymbol{b_0}$ and $\boldsymbol{b_1}$ of β_0 and β_1 are obtained, the ***fitted values*** (or ***predicted values***), denoted $\boldsymbol{\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_n}$, are defined as

Fitted (Predicted) Values:

$$\hat{\pi}_i \;=\; \frac{e^{b_0+b_1X_i}}{1+e^{b_0+b_1X_i}} \qquad \text{for } i = 1, 2, \ldots, n \tag{4}$$

The ith **fitted value** $\hat{\pi}_i$ is the **estimated probability** that an individual whose predictor value is X_i will have a response value Y equal to one.

In general, if we plug any value in for X on the right side of (4), we get the **estimated probability** that Y will equal one for that X value.

1.2 Odds, Odds Ratios, and the Interpretation of the Estimate b_1 of β_1

- The ***odds*** of an outcome is a **ratio** of **two probabilities**: the **probability** that it *will* occur divided by the **probability** that it *won't* occur*:

> **Odds**: The **odds** of a *success* ($Y = 1$) is
>
> $$\frac{\pi}{1-\pi} \;=\; \frac{P(Y=1)}{P(Y=0)} \tag{5}$$

Thus the logistic regression model (2) expresses the ***log odds*** as a **linear function** of $\boldsymbol{X}$.

* In sports, horse racing, gambling, etc. the ***odds*** of an outcome refers to the **reciprocal** of the definition given in (5).

- We can examine how the **odds** of a *success* ($Y = 1$) **changes** as $\boldsymbol{X}$ **increases** by **one unit**. The fitted value at some (generic) value of X is

$$\hat{\pi}_1 \;=\; \frac{e^{b_0+b_1X}}{1+e^{b_0+b_1X}}$$

and after increasing X by one unit, the fitted value changes to

$$\hat{\pi}_2 \;=\; \frac{e^{b_0+b_1(X+1)}}{1+e^{b_0+b_1(X+1)}}\,.$$

From

$$\log\left(\frac{\hat{\pi}_1}{1-\hat{\pi}_1}\right) \;=\; b_0+b_1X \qquad \text{and} \qquad \log\left(\frac{\hat{\pi}_2}{1-\hat{\pi}_2}\right) \;=\; b_0+b_1(X+1)$$

and using the fact that

$$\log\left(\frac{\hat{\pi}_2/(1-\hat{\pi}_2)}{\hat{\pi}_1/(1-\hat{\pi}_1)}\right) \;=\; \log\left(\frac{\hat{\pi}_2}{1-\hat{\pi}_2}\right) - \log\left(\frac{\hat{\pi}_1}{1-\hat{\pi}_1}\right)$$

it's easy to see that

$$b_1 \;=\; \log\left(\frac{\hat{\pi}_2/(1-\hat{\pi}_2)}{\hat{\pi}_1/(1-\hat{\pi}_1)}\right). \tag{6}$$

- We define the (estimated) ***odds ratio*** of an individual's response being a *success* ($Y = 1$), for a **one-unit increase** in $\boldsymbol{X}$, denoted $\widehat{\textbf{OR}}$, as:

(Estimated) Odds Ratio: The (estimated) **odds ratio** of an individual's response being a *success* ($Y = 1$), for a **one-unit increase** in $\boldsymbol{X}$, is

$$\hat{\text{OR}} \;=\; \frac{\hat{\pi}_2/(1-\hat{\pi}_2)}{\hat{\pi}_1/(1-\hat{\pi}_1)}$$

- Thus from (6),

$$b_1 \;=\; \log\left(\hat{\text{OR}}\right)$$

and so, exponentiating both sides, the (estimated) **odds ratio** can be written in terms of b_1 as

$$\hat{\text{OR}} \;=\; e^{b_1}$$

and represents the odds of *success* ($Y = 1$) at $X + 1$, relative to the odds of *success* at X.

1.3 Generalized Linear Models

- The **logistic regression model** (2) is an example of a so-called *generalized linear model*.

Recall that ***generalized linear models*** are a class of statistical models in which:

1. The response variable Y isn't necessarily normally distributed.
2. Some function $g(\mu)$ of the mean response $\mu = \text{E}(Y)$ (called the ***link function***) is expressed as a linear function of X, i.e.

$$g(\mu) \;=\; \beta_0 + \beta_1 X\,.$$

For logistic regression, $Y \sim \text{Bernoulli}(\pi)$, $\mu = \text{E}(Y) = \pi$, and $g(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$.