# Logistic Regression

(gustavoc@msu.edu)

## Binary Outcomes

Many outcomes of interest are binary, implying that they can take two values (say, 0/1). Disease is a typical example of this.

Binary random variables follow Bernoulli distributions: $p(Y_i = 1) = \theta$ or $p(Y_i = 1) = \theta$ ; $p(Y_i = 0) = 1 - \theta$, or,

$$p(Y_i = y_i|\theta) = \theta^{y_i}(1-\theta)^{1-y_i}$$

(Note: above, $Y_i$ denotes the random variable and $y_i$ represents the realized value)

## Maximum likelihood estimation of the success probability

The likelihood function is the joint probability of the data given the parameters, evaluated at the observed values of the data $S = \{Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n\}$ viewed as a function of the parameters ($\theta$). In the case of a random sample, the joint probability of the data is simply the product of the probability of each of the data points, thus

$$\begin{aligned} \mathrm{p}(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n|\theta) &= \mathrm{p}(Y_1|\theta) \times \mathrm{p}(Y_2|\theta) \times \dots \times \mathrm{p}(Y_n|\theta) \\ &= \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} \\ &= \theta^{\sum_i y_i}(1-\theta)^{\sum_i 1-y_i} = \theta^{n\bar{y}}(1-\theta)^{n(1-\bar{y})} \end{aligned}$$

Thus, the likelihood function is

$$L(\theta|\, y_1, \dots, y_n) = \theta^{n\bar{y}}(1-\theta)^{n(1-\bar{y})}$$

The Maximum Likelihood estimator (MLE) is obtained by maximizing $L(\theta|\, y_1, \dots, y_n)$ with respect to $\theta$; the same estimate can be obtained by maximizing the log-likelihood

$$l(\theta|\, y_1, \dots, y_n) = \log\{L(\theta|\, y_1, \dots, y_n)\} = n\bar{y}\log(\theta) + n(1-\bar{y})\log(1-\theta)$$

Differentiating with respect to $\theta$ we get

$$\frac{dl(\theta|\, y_1, \dots, y_n)}{d\theta} = \frac{n\bar{y}}{\theta} - \frac{n(1-\bar{y})}{(1-\theta)}$$

Setting the derivative equal to zero we get maximum the MLE

$$\frac{n\bar{y}}{\hat{\theta}} = \frac{n(1-\bar{y})}{(1-\hat{\theta})}$$
$$\frac{\bar{y}}{(1-\bar{y})} = \frac{\hat{\theta}}{(1-\hat{\theta})} \text{ (assuming } \bar{y} \neq 1)$$
$$\hat{\theta} = \bar{y}$$

Thus, the MLE of the success probability is simply the sample mean of the data, which is not surprising considering that $E[Y_i] = \theta$!

**Logistic Regression**

We are often interested on learning the effects of some factors (e.g., sex) and covariates (e.g., age) on the probability of a binary outcome ($\theta$, e.g., a disease probability). In the previous example this probability was assumed to be the same for all individuals. To model effects of covariates on $\theta$, in logistic regression, we make $\theta$ a function of covariates.

Since $\theta \in [0,1]$ we cannot model $\theta$ directly using linear regression because a linear function can take any value in the real line. To deal with this problem we introduce a “link” function (e.g., probit, logit). A link function maps from the real line onto the $[0,1]$. The most commonly used link is the logit which is the logarithm of the odds of success, that is: $log\left(\frac{\theta_i}{1-\theta_i}\right)$. This function can take values in the real line, thus, we can model the logit using linear methods

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \mu + X_{i1}\beta_1 + \cdots + X_{ip}\beta_p. \qquad [1]$$

Note that the above regression is a regression for the probability, not for the data, thus, it typically does not include an error term (in some over-dispersed models it may contain an error).

**From regression to probabilities**

Solving [1] for $\theta_i$ gives

$$\theta_i = \frac{\exp\{\mu + X_{i1}\beta_1 + \cdots + X_{ip}\beta_p\}}{1 + \exp\{\mu + X_{i1}\beta_1 + \cdots + X_{ip}\beta_p\}}. \qquad [2]$$

Letting the right-hand side of [1], i.e., the regression function, be $\eta_i = \mu + X_{i1}\beta_1 + \cdots + X_{ip}\beta_p$ then we have: $\theta_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$.

## Likelihood function for the logistic regression model

The likelihood function is the probability of the data given the parameters. As before, we will assume conditional independence, meaning that

$$\mathrm{p}\left(Y_1, Y_2, \dots, Y_n | \mu, \beta_1, \dots, \beta_p, X\right) = \mathrm{p}\left(Y_1 | \mu, \beta_1, \dots, \beta_p, X\right) \times \mathrm{p}\left(Y_2 | \mu, \beta_1, \dots, \beta_p, X\right) \times \dots \times \mathrm{p}\left(Y_n | \mu, \beta_1, \dots, \beta_p, X\right)$$

The probability of the *i^th^* data-point is:

$$p(Y_i = 1) = \theta_i = \frac{e^{\eta_i}}{1 - e^{\eta_i}} \; ; \; p(Y_i == 0) = 1 - \theta_i = 1 - \frac{e^{\eta_i}}{1 - e^{\eta_i}} = \frac{1}{1 + e^{\eta_i}}$$

or, $p(Y_i = y_i) = \left[\frac{e^{\eta_i}}{1 - e^{\eta_i}}\right]^{y_i} \left[\frac{1}{1 - e^{\eta_i}}\right]^{1 - y_i}$

Therefore, assuming conditional independence, the joint likelihood becomes

**Likelihood**: $L\left(\mu, \beta_1, \dots, \beta_p | Y_1 = y_1, \dots, Y_n = y_n\right) = \prod_{i=1}^{n} \left[\frac{e^{\eta_i}}{1 - e^{\eta_i}}\right]^{y_i} \left[\frac{1}{1 - e^{\eta_i}}\right]^{1 - y_i}$ [3]

Note that above: (i) $y_i$ is a realized value of the corresponding Bernoulli random variable ($Y_i$), therefore, $y_i$ can take values either 0 or 1. (ii) $\eta_i = \mu + X_{i1}\beta_1 + \cdots + X_{ip}\beta_p$ is a function of both covariates ($X_{ij}, j = 1, \dots, p$) and parameters ($\mu$, $\beta_j$).

Therefore, the log-likelihood function is

$$l\left(\mu, \beta_1, \dots, \beta_p | y_1, y_2, \dots, y_n\right) = \sum_{i=1}^{n} y_i \log(\theta_i) + (1 - y_i) \log(1 - \theta_i) \qquad [4]$$

where $\theta_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$ and $1 - \theta_i = \frac{1}{1 + e^{\eta_i}}$.

The entry logisticRegression.md in our gitHub repository implements this function in R.

## Maximum Likelihood estimation

Maximum likelihood estimates are obtained by maximizing [4] with respect to the parameters $(\mu, \beta_1, \dots, \beta_p)$. The function glm in R fits logistic regression via maximum likelihood. We can also fit a logistic regression using a general-purpose optimization algorithm (e.g., optim in R). The entry logisticRegression.md in in our gitHub repository shows how to fit logistic regression using glm and optim.

## Bayesian Model

A Bayesian model is defined by the likelihood function and the prior. The likelihood function for the logistic regression is given by [3]. The unknown parameters are the regression coefficients, denoted hereinafter by $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ [note, while above we made distinction of the intercept, hereinafter, for ease of presentation we assume that the first regression corresponds to the intercept]. Regression coefficients can take any values in the real line. Thus, a common choice for the prior is the normal distribution. The most common approach is to use IID normal priors with zero mean and large variance, this gives a relatively flat prior and a posterior distribution that it is dominated by the likelihood. Adopting these assumptions, we have

**Prior**: $$p(\boldsymbol{\beta}) = \prod_{j=1}^{p} \frac{e^{-\frac{b_j^2}{2\sigma_b^2}}}{\sqrt{2\pi\sigma_b^2}} \qquad [5]$$

According to Bayes' theorem, the posterior distribution is proportional to the product of the likelihood [3] times the prior [5]:

$$p(\boldsymbol{\beta}|y_1, \dots, y_n) = \frac{p(y_1, \dots, y_n|\boldsymbol{\beta})p(\boldsymbol{\beta})}{p(y_1, \dots, y_n)} \propto p(y_1, \dots, y_n|\boldsymbol{\beta})p(\boldsymbol{\beta})$$

Therefore, using [3] and [5] we have

$$p(\boldsymbol{\beta}|y_1, \dots, y_n) \propto \left\{\prod_{i=1}^{n} \left[\frac{e^{\eta_i}}{1-e^{\eta_i}}\right]^{y_i} \left[\frac{1}{1-e^{\eta_i}}\right]^{1-y_i}\right\} \times \left\{\prod_{j=1}^{p} \frac{e^{-\frac{b_j^2}{2\sigma_b^2}}}{\sqrt{2\pi\sigma_b^2}}\right\}. \quad [6]$$

The above distribution does not have a closed form. Therefore, we need to use Monte Carlo methods: draw samples from the posterior distribution and use these samples to approximate the quantities of interest (e.g., posterior means, posterior standard deviations, credibility regions, etc.).

In the case of the linear model we used a Gibbs sampler to draw samples from the posterior distribution. However, implementing a Gibbs sampler requires sampling from the fully-conditional distributions. In the logistic regression the fully conditional distributions don't have a closed form; thus, we cannot use a Gibbs sampler. Instead we will use a Metropolis Hastings algorithm.