# Logistic Regression and Generalized Linear Model

Michael

School of Mathematics and Statistics, UCD

School of Economics, University of Nottingham

## 1 Logistic Regression

Assume the probability $\pi_i$ of $y_i$ happened depends on a vector of observed covariates $x_i$ (e.g., education, income). The simples idea would be to let $\pi_i$ be a linear function of the covariates, say

$$\pi_i = x_i'\beta \tag{1.1}$$

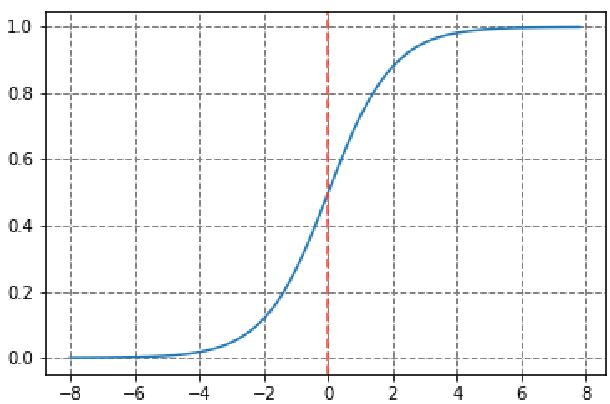
where $\beta$ is a vector of regression coefficients. The left hand side of equation (1.1) has to be between 0 and 1, but the linear predictor $x_i'\beta$ on the right hand side can take any real value. To solve this problem, we transform the probability into odds ratio:

$$\text{odds} = \frac{\pi}{1-\pi}$$
$$\eta = \log(\text{odds}) = \log(\frac{\pi_i}{1-\pi_i})$$
$$\pi_i = \frac{e^\eta}{1+e^\eta}$$

In the end, we can get a continuous function between covariates $x$ and dependent variable probability $\pi$:

$$\pi = h_\theta(x) = \frac{1}{1+e^{-x'\theta}} = \frac{1}{1+e^{-z}}; \quad z = x'\theta \tag{1.2}$$

Equation (1.2) is called **sigmoid function** in machine learning.



### 1.1 Binary Case

For binary case, let us assume that

$$\mathbb{P}(y = 1|x; \theta) = h_\theta(x)$$
$$\mathbb{P}(y = 0|x; \theta) = 1 - \theta_\theta(x)$$

Note this can be written more compactly as

$$P(y|x; \theta) = [h_\theta(x)]^y [1 - h_\theta(x)]^{1-y}$$

The loglikelihood function becomes

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{N} \mathbb{P}(y^i|x^i; \theta) \\ &= \prod_{i=1}^{N} [h_\theta(x)]^y [1 - h_\theta(x)]^{1-y} \end{aligned}$$

The gradient descent algorithm have the following update rule:

$$\theta = \theta + \alpha \Delta_\theta L(\theta)$$

where $\Delta_\theta L(\theta)$ is just the derivative of our loglikelihood function, which is

$$\begin{aligned} \frac{\partial L(\theta)}{\partial \theta} &= [y - h_\theta(x)] x_j \\ &= [y - \frac{1}{1 + e^{-x'\theta}}] x_j \end{aligned}$$

Although we could use gradient descent algorithm to estimate coefficients. It turns out that gradient descent is not efficient. Hence, people turned to Newton-Raphson algorithm for maximising our loglikelihood function:

- When the derivative $\frac{\partial L(\theta)}{\partial \theta} = 0$, one have the optimal value of $L(\theta)$;
- We use newton's method to solve $\frac{\partial L(\theta)}{\partial \theta} = 0$
- Hence, we need the second derivative of $L(\theta)$ or the first derivative of $\frac{\partial L(\theta)}{\partial \theta}$

The Hessian can be derived as follows[1]:

$$\begin{aligned} \frac{\partial L^2}{\partial \theta^2} &= -\frac{x_j' e^{-x'\theta}}{(1 + e^{-x'\theta})^2} x_j \\ &= -X^T \text{diag}[h_\theta(x)(1 - h_\theta(x)] X \\ &= -\sum_{i=1}^{N} x_i x_i^T [h_\theta(x_i)(1 - h_\theta(x_i)] \end{aligned} \tag{1.3}$$

According to Newton's rule and equation (1.3), we can have the updating rule:

$$\theta^{new} = \theta^{old} - (\frac{\partial^2 L(\theta)}{\partial \theta \partial \theta^T})^{-1} \frac{\partial L(\theta)}{\partial \theta}$$

where the derivatives are evaluated at $\theta^{old}$.

[1] The detailed derivation can be found in the lecture by Jia LI: `http://personal.psu.edu/jol2/course/stat597e/notes2/logit.pdf`

## 2 Multinomial Model

When we $K$ classes, the model has the form

$$\begin{aligned}
\log \frac{\mathbb{P}(G=1|X=x)}{\mathbb{P}(G=K|X=x)} &= \beta_{10} + \beta_1^T x \\
\log \frac{\mathbb{P}(G=2|X=x)}{\mathbb{P}(G=K|X=x)} &= \beta_{20} + \beta_2^T x \\
&\vdots \\
\log \frac{\mathbb{P}(G=K-1|X=x)}{\mathbb{P}(G=K|X=x)} &= \beta_{(K-1)0} + \beta_{K-1}^T x
\end{aligned}$$

Although the model uses the last class as the denominator in the odds-ratios, the choice of denominator is arbitrary in that the estimates are equivalent under this choice. A simple calculation shows that

$$\begin{aligned}
\mathbb{P}(G=k|X=x) &= \frac{\exp(\beta_{k0} + \beta_k^T x)}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \beta_l^T x)}, \quad k = 1, \cdots, K-1, \\
\mathbb{P}(G=K|X=x) &= \frac{1}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \beta_l^T x}
\end{aligned}$$

## 3 Generalized Linear Models

One can read the following textbook to understand generalized linear models well:

- An Introduction to Generalized Linear Models, by *Annette J. Dobson, Adrian G Barnett*