# Lecture Notes - 04: Logistic Regression

Dihui Lai

March 29, 2020

## Contents

# 1 Likelihood Function

## 1.1 Definition

If a set of random variables $Y_1$, $Y_2$ ... $Y_n$ has a joint probability distribution density/mass $f(y_1, y_2, ...y_n; \theta)$, where $\theta$ is a set of parameters, the likelihood function is defined as

$$L(\theta) = f(y_1, y_2, ...y_n; \theta) \tag{1}$$

## 1.2 Example: Bernoulli/Binomial Distribution

Assuming an event has two possible outcomes $y = 1$ or $y = 0$, with probability $p$ of being 1, i.e. the outcome follows a Bernoulli distribution. As we learned in lecture 2, the probability mass function is

$$f(y;p) = \begin{cases} p, & y = 1 \\ 1-p, & y = 0 \end{cases}$$

Or

$$f(y;p) = p^y(1-p)^y$$

The probability mass distribution (or the likelihood function by definition) for $n$ independent events is

$$L(p_1, p_2, ...p_n) = f(y_1, y_2, ...y_n; p_1, p_2, ..., p_n) = \prod_{i=1}^{n} p_i^{y^i}(1-p_i)^{1-y^i}$$

To interpreting the likelihood function, let us consider the underlying parameters are the same i.e. $p = p_1 = p_2 ... = p_n$ for all the data entries observed. And we have the likelihood function as

$$L(p) = \prod_{i=1}^{n} p^{y^i}(1-p)^{1-y^i}$$

Let us consider the following cases $n = 1$, $n = 2$ and any $n$. What kind of $p$ that can maximize the likelihood function $L(p)$?

- $n = 1$ (1 observation): The likelihood function is $L(p) = p^y(1-p)^{1-y}$.

| Observations | $L(p)$ | $L_{max}(p)$ |
|---|---|---|
| $y = 0$ | $L(p) = 1-p$ | $L_{max} = 1$ at $p = 0$ |
| $y = 1$ | $L(p) = p$ | $L_{max} = 1$ at $p = 1$ |

- $n = 2$ (2 observations): The likelihood function is $L(p) = p^{y^1+y^2}(1-p)^{(1-y^1)+(1-y2)}$. Given the

| Observations | $L(p)$ | $L_{max}(p)$ |
|---|---|---|
| $y^1 = 0, y^2 = 0$ | $L(p) = (1-p)^2$ | $L_{max} = 1$ at $p = 0$ |
| $y^1 = 1, y^2 = 1$ | $L(p) = p^2$ | $L_{max} = 1$ at $p = 0$ |
| $y^1 = 0, y^2 = 1$ | $L(p) = p(1-p)$ | $L_{max} = 0.25$ at $p = 0.5$ |

- $n = n_1 + n_0$ ($n$ observations with $n_1$ 1s and $n_0$ 0s): The likelihood function is $L(p) = p^{n_1}(1-p)^{n_0}$. The likelihood function is maximized when

$$\frac{\partial \ell}{\partial p} = 0, \text{ where } \ell = \log(L(p)) = n_1 \log(p) + n_0 \log(1-p) \tag{2}$$

  Solve equation (3) for $p$, we have

$$\begin{aligned} \frac{\partial \ell}{\partial p} = \frac{n_1}{p} - \frac{n_0}{1-p} = 0 \\ \Rightarrow n_1 - n_1 p - n_0 p = 0 \\ \Rightarrow p = \frac{n_1}{n_1 + n_0} \end{aligned}$$

Overall, $p$ maximize the likelihood function when it takes the value of the mean of observed $y$s

## 1.3 Example: Multinomial Distribution

The multinomial distribution has density function

$$f(x_1, x_2, x_3, ...x_c) = \frac{N!}{x_1!x_2!...x_c!} p_1^{x_1} p_2^{x_2} ... p_c^{x_c}$$

If we perform one experiemnts (N=1), the likelihood function is accordingly, $L_i = \prod_{j=1}^{c} p_j^{x_j^i}$ the likelihood function of n experiments is then $L = \prod_{i=1}^{n} \prod_{j=1}^{c} p_j^{x_j^i}$. Note $x_j^i$ is either 1 or 0

The log-likelihood function (a.k.a log-loss) is

$$\ell = \log(L) = \sum_{i=1}^{n} \sum_{j=1}^{c} x_j^i \log(p_j)$$

The $p_k$ that maximize the log-likelihood function that subject to the constraint $\sum_{j=1}^{c} p_j = 1$ has to satisfy the following condition

$$\begin{cases} \dfrac{\partial}{\partial p_k} \left( \ell - \lambda \sum_{i=1}^{n} (1 - \sum_{j=1}^{c} p_j) \right) = 0 \\ \dfrac{\partial}{\partial \lambda} \left( \ell - \lambda \sum_{i=1}^{n} (1 - \sum_{j=1}^{c} p_j) \right) = 0 \end{cases}$$

$$\Rightarrow \begin{cases} \sum_{i=1}^{n} \frac{\partial}{\partial p_k} \left( \sum_{j=1}^{c} x_j^i \log(p_j) - \lambda(1 - \sum_{j=1}^{c} p_j) \right) = 0 \\ \sum_{i=1}^{n} \frac{\partial}{\partial \lambda} \left( \sum_{j=1}^{c} x_j^i \log(p_j) - \lambda(1 - \sum_{j=1}^{c} p_j) \right) = 0 \end{cases}$$

$$\Rightarrow \lambda = -\frac{x_k}{p_k}$$

where $x_k$ is the total number of outcome $k$. Because

$$\sum_{k=1}^{c} x_k = n$$

We have

$$-\sum_{k=1}^{c} \lambda p_k = n \Rightarrow \lambda = -n$$

Therefore

$$p_k = \frac{x_k}{n} \tag{3}$$

# 2 Logistic Regression

## 2.1 Likelihood Function

In general, every events could have its own underlying parameter $p$. For n-independent events, let us assume the parameters are $p_1$, $p_2$, ..., $p_n$ respectively. The corresponding log-likelihood function is thus

$$\ell(p_1, p_2, ...p_n) = \sum_{i=1}^{n} \left( y^i \log(p_i) + (1 - y^i) \log(1 - p_i) \right) \tag{4}$$

The log-likelihood function is the defined as the log transformation of the likelihood function

$$\ell = \log(L) = \sum_{i=1}^{n} y^i \log(p_i) + (1 - y^i) \log(1 - p_i) \tag{5}$$

## 2.2 Parameter Model

The parameter $p_i$ is modeled as a logistic function of a set of $m$ predictors $x_1^i$, $x_2^i$, ...$x_m^i$ or $\vec{x}^i$ in vector notation.

$$p_i = \frac{1}{1+\exp(-\vec{\beta}\cdot\vec{x}^i)} \tag{6}$$

## 2.3 Maximum Likelihood Estimation

The optimal model chooses $\beta$s that maximize the likelihood function $\ell$, at the optimal point $\beta$s satisfy the following equations.

$$\frac{\partial \ell}{\partial \beta_j} = 0 \ , \ j = 1, 2, ..., m \tag{7}$$

Use $\ell$'s definition in equation (4) and formula (5), we have

$$\begin{aligned} \ell &= \sum_{i=1}^{n} y^i \log \frac{p_i}{1-p_i} + \log(1-p_i) \\ &= \sum_{i=1}^{n} y^i(\vec{\beta}\cdot\vec{x}^i) - \log(1+\exp(\vec{\beta}\cdot\vec{x}^i))) \end{aligned}$$

Insert it into equation (6), we have

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \left( y^i - \frac{1}{1+\exp(-\vec{\beta}\cdot\vec{x}^i)} \right) x_j^i = 0 \ , \ j = 1, 2, 3, ..., m$$

To get the optimal $\beta$s, we need to solve the equation set. However, it is hard to do analytically, because of the nonlinear terms that contain $\beta$ $\frac{1}{1+\exp(-\vec{\beta}\cdot\vec{x}^i)}$. However, we can solve the problem numerically, using Newton-Raphson method.

# 3 Newton-Raphson Method

## 3.1 Single Variable

Consider a log-likelihood function of one parameter $\ell(\beta)$. In general, $\ell$ can be of any function and complex. With the hope that its derivative $\ell'$ is simpler, we use Taylor expansion for approximation around some point $\beta_0$

$$\ell(\beta) \sim \ell(\beta_0) + \ell'(\beta_0)(\beta-\beta_0) + \frac{1}{2}\ell''(\beta_0)(\beta-\beta_0)^2 \tag{8}$$

The derivative of equation (7) is thus

$$\ell'(\beta) \sim 0 + \ell'(\beta_0) + \ell''(\beta_0)(\beta - \beta_0) \tag{9}$$

The $\beta^*$ that minimizes the log-likelihood function have $\ell'(\beta) = 0$ at the point i.e. $\ell'(\beta)|_{\beta=\beta^*} = 0$. Using equation (8), we have

$$\ell'(\beta_0) + \ell''(\beta_0)(\beta^* - \beta_0) = 0 \tag{10}$$

$$\Rightarrow \beta^* = \beta_0 - \frac{\ell'(\beta_0)}{\ell''(\beta_0)} \tag{11}$$

Recall that this is only an approximation solution and $\beta^*$ is not exactly the optimal point with an arbitrarily chosen $\beta_0$. However, we can hope that equation (10) brings us a little closer to the optimal point. To get a more accurate solution, we will need to use equation (10) iteratively i.e.

$$\beta_{k+1} = \beta_k - \frac{\ell'(\beta_k)}{\ell''(\beta_k)}, \text{ until } |\beta_{k+1} - \beta_k| < \delta$$

Here, $|\beta_{k+1} - \beta_k| < \delta$ is the convergence condition and $\delta$ is tolerance level. $\delta$ is usually set as a small number. The algorithms says that we can stop the iteration if we are very close to the optimal point.

## 3.2 Multiple Variable

In the case where the log-likelihood function is dependent on multiple parameters $\ell(\beta)$, the Taylor expansion is

$$\ell(\beta) \sim \ell(\beta_0) + \nabla\ell(\beta_0)^T(\beta - \beta_0) + \frac{1}{2}(\beta - \beta_0)^T\mathbf{H}(\beta_0)(\beta - \beta_0) \tag{12}$$

Here $\beta$ is a $m \times 1$ column matrix and $\mathbf{H}$ is the $m \times m$ Hessian matrix, defined as

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}, \mathbf{H} = \begin{bmatrix} \frac{\partial^2 \ell}{\partial \beta_1^2} & \frac{\partial^2 \ell}{\partial \beta_1 \, \partial \beta_2} & \cdots & \frac{\partial^2 \ell}{\partial \beta_1 \, \partial \beta_m} \\ \frac{\partial^2 \ell}{\partial \beta_2 \, \partial \beta_1} & \frac{\partial^2 \ell}{\partial \beta_2^2} & \cdots & \frac{\partial^2 \ell}{\partial \beta_2 \, \partial \beta_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \ell}{\partial \beta_m \, \partial \beta_1} & \frac{\partial^2 \ell}{\partial \beta_m \, \partial \beta_2} & \cdots & \frac{\partial^2 \ell}{\partial \beta_m^2} \end{bmatrix}$$

Apply the gradient against $\beta$ on equation (11), we have

$$\nabla\ell = \nabla\ell(\beta_0) + \mathbf{H}(\beta - \beta_0) \text{ see Appendix.}$$

At the optimal point we want to have $\nabla \ell = 0$ i.e.

$$\begin{aligned}
&\nabla \ell(\beta_0) + \mathbf{H}(\beta - \beta_0) = 0 \\
\Rightarrow &\mathbf{H^{-1}} \nabla \ell(\beta_0) + (\beta - \beta_0) = 0 \\
\Rightarrow &\beta = \beta_0 - \mathbf{H^{-1}} \nabla \ell(\beta_0)
\end{aligned}$$

The Newton-Raphson algorithm for multivariate model is therefore

$$\beta_{k+1} = \beta_k - \mathbf{H^{-1}} \nabla \ell(\beta_k), \text{ until } |\beta_{k+1} - \beta_k| < \delta \tag{13}$$

# 4 Appendix

## 4.1 The Gradient of Equation (11)

Starting with equation

$$\ell(\beta) = \ell(\beta_0) + \nabla \ell(\beta_0)^T (\beta - \beta_0) + \frac{1}{2} (\beta - \beta_0)^T \mathbf{H}(\beta_0)(\beta - \beta_0) \tag{14}$$

To simplify the equation, we introduce the notation $\Delta\beta = \beta - \beta_0$. It is easy to see the derivative of each element of $\Delta\beta$ against $\beta_j$ has the following property

$$\frac{\partial}{\partial \beta_j} \Delta\beta_i = \delta_{ij} \tag{15}$$

Here, $\delta_{ij}$ is the Kronecker delta, having the property $\delta_{ij} = \begin{cases} 0 & \text{if } i \neq j, \\ 1 & \text{if } i = j. \end{cases}$

On the other hand, if we write the log likelihood using the elements in the matrices, we have

$$\ell(\beta) = \ell(\beta_0) + \sum_{a=1}^{m} \frac{\partial \ell(\beta_0)}{\partial \beta_a} (\Delta\beta_a) + \sum_{a,b=1}^{m} \frac{1}{2} \Delta\beta_a H_{ab}(\beta_0) \Delta\beta_b \tag{16}$$

Let us look at each term on the R.H.S of the equation when we take the partial derivative of $\ell$ against $\beta_j$.

- The first term becomes 0 as it is constant $\nabla \ell(\beta_0) = 0$.

- In the second term, only $\Delta\beta_a$ is dependent on $\beta$ and we have

$$\begin{aligned}
&\frac{\partial}{\partial\beta_j}\left(\sum_{a=1}^{m}\frac{\partial\ell(\beta_0)}{\partial\beta_a}(\Delta\beta_a)\right)\\
&=\sum_{a=1}^{m}\frac{\partial\ell(\beta_0)}{\partial\beta_a}\frac{\partial\Delta\beta_a}{\partial\beta_j}\\
&=\sum_{a=1}^{m}\frac{\partial\ell(\beta_0)}{\partial\beta_a}\delta_{aj}\\
&=\frac{\partial\ell(\beta_0)}{\partial\beta_j}
\end{aligned}$$

In matrix format, we have

$$\nabla\left(\sum_{a=1}^{m}\frac{\partial\ell(\beta_0)}{\partial\beta_a}(\Delta\beta_a)\right)=\nabla\ell(\beta_0)$$

- the third term has two variables dependent on $\beta$ $\Delta\beta_a$ and $\Delta\beta_b$

$$\begin{aligned}
&\frac{\partial}{\partial\beta_j}\left(\sum_{a,b=1}^{m}\frac{1}{2}\Delta\beta_a H_{ab}(\beta_0)\Delta\beta_b\right)\\
&=\sum_{a,b=1}^{m}\frac{1}{2}\delta_{aj}H_{ab}(\beta_0)\Delta\beta_b+\sum_{a,b=1}^{m}\frac{1}{2}\Delta\beta_a H_{ab}(\beta_0)\delta_{bj}\\
&=\sum_{b=1}^{m}\frac{1}{2}H_{jb}(\beta_0)\Delta\beta_b+\sum_{a=1}^{m}\frac{1}{2}\Delta\beta_a H_{aj}(\beta_0)\\
&=\sum_{b=1}^{m}\frac{1}{2}H_{jb}(\beta_0)\Delta\beta_b+\sum_{a=1}^{m}\frac{1}{2}H_{ja}(\beta_0)\Delta\beta_a,\text{ use the fact that } H_{aj}=H_{ja}\\
&=\sum_{d=1}^{m}H_{jd}(\beta_0)\Delta\beta_c,\ \text{(a, b are dummy indices, set them to be c)}
\end{aligned}$$

In matrix format we have

$$\nabla\left(\sum_{a,b=1}^{m}\frac{1}{2}\Delta\beta_a H_{ab}(\beta_0)\Delta\beta_b\right)=\mathbf{H}\Delta\beta$$

Therefore we have

$$\nabla\ell(\beta)=\nabla\ell(\beta_0)+\nabla\left(\sum_{a=1}^{m}\frac{\partial\ell(\beta_0)}{\partial\beta_a}(\Delta\beta_a)\right)+\nabla\left(\sum_{a,b=1}^{m}\frac{1}{2}\Delta\beta_a H_{ab}(\beta_0)\Delta\beta_b\right) \tag{17}$$

$$=0+\nabla\ell(\beta_0)+\mathbf{H}\Delta\beta \tag{18}$$