Logistic Regression Maximum Likelihood Parameter Estimation Procedure Derivation

1. Preliminaries

- Sample size: m
- Target variable: $\begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix} = \vec{y}$, where $y^{(i)} \in \{0,1\} \rightarrow Y \sim Bernoulli(p)$, where $p = \Pr\{Y = 1\}$
- Input features: $X = [\vec{x}^{(1)} \quad \vec{x}^{(2)} \quad \dots \quad \vec{x}^{(m)}]$, where $\vec{x}^{(i)} = \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_q^{(i)} \end{bmatrix}$, for $i \in \{1,2, \dots, m\}$
- Goal: Reliably estimate Y given $\vec{x}$

2. Method

2.1 Logistic Regression Set-up

We will use logistic regression on p and estimate the parameters via the maximum likelihood method.

Mathematically, for a set of features $\vec{x} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_q \end{bmatrix}$ (we added an extra constant feature to the original vector $\vec{x}$ for notation elegance purposes, and define $x_0 = 1$) and parameters $\vec{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_q \end{bmatrix}$, we assume the estimate of p, denoted as $\hat{p}(\vec{x}^T\vec{\beta})$ is calculated as follows:

$$\operatorname{logit}\left(\hat{p}(\vec{x}^T\vec{\beta})\right) = \ln\left(\frac{\hat{p}(\vec{x}^T\vec{\beta})}{1 - \hat{p}\widehat{(\vec{x}^T\vec{\beta})}}\right) = \vec{x}^T\vec{\beta} \Leftrightarrow \hat{p}(\vec{x}^T\vec{\beta}) = \frac{1}{1 + e^{-\vec{x}^T\vec{\beta}}}$$

2.2 Maximum Likelihood Expressions

We use the maximum likelihood method:

Let $\mathcal{L}\left(\vec{\beta}|(X, \vec{y})\right) = \Pr\left\{Y^{(1)} = y^{(1)}, Y^{(2)} = y^{(2)}, \dots, Y^{(m)} = y^{(m)}\right\}$

Assuming an independently and identically distributed sample, we have the following expressions:

$$
\begin{aligned}
\mathcal{L}\left(\beta_0,\vec{\beta}|(X,\vec{y})\right) &= \prod_{i=1}^{m}\Pr\{Y=y^{(i)}\} = \prod_{i=1}^{m}\Pr\{Y=1\}^{y^{(i)}} \times (1-\Pr\{Y=1\})^{1-y^{(i)}} \\
&= \prod_{i=1}^{m}\left(\hat{p}\left(\vec{x}^{\,(i)T}\vec{\beta}\right)\right)^{y^{(i)}} \times \left(1-\hat{p}\left(\vec{x}^{\,(i)T}\vec{\beta}\right)\right)^{1-y^{(i)}} \\
&= \prod_{i=1}^{m}\left(\frac{1}{1+e^{-\vec{x}^{(i)T}\vec{\beta}}}\right)^{y^{(i)}} \times \left(1-\frac{1}{1+e^{-\vec{x}^{(i)T}\vec{\beta}}}\right)^{1-y^{(i)}} \\
&= \prod_{i=1}^{m}\left(\frac{1}{1+e^{-\vec{x}^{(i)T}\vec{\beta}}}\right)^{y^{(i)}} \times \left(\frac{e^{-\vec{x}^{(i)T}\vec{\beta}}}{1+e^{-\vec{x}^{(i)T}\vec{\beta}}}\right)^{1-y^{(i)}} \\
&= \prod_{i=1}^{m}\left(\frac{e^{\vec{x}^{(i)T}\vec{\beta}}}{1+e^{\vec{x}^{T}\vec{\beta}}}\right)^{y^{(i)}} \times \left(\frac{1}{1+e^{\vec{x}^{(i)T}\vec{\beta}}}\right)^{1-y^{(i)}} \\
&= \prod_{i=1}^{m}\frac{e^{y^{(i)}\vec{x}^{(i)T}\vec{\beta}}}{1+e^{\vec{x}^{(i)T}\vec{\beta}}}
\end{aligned}
$$

The goal is to maximize this function. This implies that if we maximize the natural logarithm of this function (which is easier to handle), we will achieve the same goal (since $\ln\left(f(.)\right)$ moves in the same direction as $f(.)$, for a function $f(.) > 0$).

Let $\ell\left(\vec{\beta}|\vec{x}\right) = \ln\left(\mathcal{L}\left(\beta_0,\vec{\beta}|(\vec{x},\vec{y})\right)\right)$. Thus, we have the following expression:

$$\ell\left(\vec{\beta}|\vec{x}\right) = \sum_{i=1}^{m} y^{(i)}\vec{x}^{(i)T}\vec{\beta} - \ln\left(1+e^{\vec{x}^{(i)T}\vec{\beta}}\right)$$

To obtain an optimal point, we can calculate the partial derivative of the log-likelihood function with respect to the parameter of interest and setting the derivative to zero:

For $j \in \{0,1,\dots,q\}$,

$$
\begin{aligned}
\frac{\partial\ell\left(\vec{\beta}|\vec{x}\right)}{\partial\beta_j} &= \sum_{i=1}^{m} y^{(i)}x_j^{(i)} - \frac{x_j^{(i)}e^{\vec{x}^{(i)T}\vec{\beta}}}{1+e^{\vec{x}^{(i)T}\vec{\beta}}} = \sum_{i=1}^{m}\left(y^{(i)} - \frac{e^{\vec{x}^{(i)T}\vec{\beta}}}{1+e^{\vec{x}^{(i)T}\vec{\beta}}}\right)x_j^{(i)} \\
&= \sum_{i=1}^{m}\left(y^{(i)} - \frac{1}{1+e^{-\vec{x}^{(i)T}\vec{\beta}}}\right)x_j^{(i)}
\end{aligned}
$$

Thus,$\ell\left(\vec{\beta}|\vec{x}\right)$ is at an optimal value when varying β_j such that $\sum_{i=1}^{m}\left(y^{(i)} - \frac{1}{1+e^{-\vec{x}^{(i)T}\vec{\beta}}}\right)x_j^{(i)} = 0$.

If we show that $\frac{\partial \ell^2(\vec{\beta}|\vec{x})}{\partial \beta_j^2} < 0$, we will establish that β_j's value that yields $\frac{\partial \ell(\vec{\beta}|\vec{x})}{\partial \beta_j} = 0$ maximizes $\ell(\vec{\beta}|\vec{x})$. For $j \in \{0,1,\dots,q\}$,

$$\frac{\partial \ell^2(\vec{\beta}|\vec{x})}{\partial \beta_j^2} = \frac{\partial}{\partial \beta_j}\left(\sum_{i=1}^{m}\left(y^{(i)} - \frac{1}{1+e^{-\vec{x}^{(i)T}\vec{\beta}}}\right)x_j^{(i)}\right)$$

$$= \sum_{i=1}^{m}\left(-\frac{-1}{\left(1+e^{-\vec{x}^{(i)T}\vec{\beta}}\right)^2} \times -x_j^{(i)} e^{-\vec{x}^{(i)T}\vec{\beta}}\right)x_j^{(i)}$$

$$= -\sum_{i=1}^{m}\left(\frac{e^{-\vec{x}^{(i)T}\vec{\beta}}}{\left(1+e^{-\vec{x}^{(i)T}\vec{\beta}}\right)^2}\right)x_j^{(i)2} < 0$$

2.3 Procedure & Formulas to Estimate the Betas

We will use the multi-variate Newton-Raphson procedure (we add a superscript $[n]$ to the variables and functions to denote that we are at iteration n):

$n = 0$: Set $\vec{\beta}^{[0]}$ to random values.

$n > 0$: $\vec{\beta}^{[n]} = \vec{\beta}^{[n-1]} - \left[\frac{\partial \ell^2(\vec{\beta}^{[n-1]}|\vec{x})}{\partial \vec{\beta}^{[n-1]^2}}\right]^{-1} \times \frac{\partial \ell(\vec{\beta}^{[n-1]}|\vec{x})}{\partial \vec{\beta}^{[n-1]}}$

We derive $\frac{\partial \ell(\vec{\beta}^{[n-1]}|\vec{x})}{\partial \vec{\beta}^{[n-1]}}$ and $\frac{\partial \ell^2(\vec{\beta}^{[n-1]}|\vec{x})}{\partial \vec{\beta}^{[n-1]^2}}$ explicitly:

$$\frac{\partial \ell(\vec{\beta}^{[n-1]}|\vec{x})}{\partial \vec{\beta}^{[n-1]}} = \begin{bmatrix} \frac{\partial \ell(\vec{\beta}|\vec{x})}{\partial \beta_0} \\ \frac{\partial \ell(\vec{\beta}|\vec{x})}{\partial \beta_1} \\ \vdots \\ \frac{\partial \ell(\vec{\beta}|\vec{x})}{\partial \beta_q} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{m}\left(y^{(i)} - \frac{1}{1+e^{-\vec{x}^{(i)T}\vec{\beta}}}\right)x_0^{(i)} \\ \sum_{i=1}^{m}\left(y^{(i)} - \frac{1}{1+e^{-\vec{x}^{(i)T}\vec{\beta}}}\right)x_1^{(i)} \\ \vdots \\ \sum_{i=1}^{m}\left(y^{(i)} - \frac{1}{1+e^{-\vec{x}^{(i)T}\vec{\beta}}}\right)x_q^{(i)} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{m}\left(y^{(i)} - \hat{p}\left(\vec{x}^{(i)T}\vec{\beta}^{[n-1]}\right)\right)x_0^{(i)} \\ \sum_{i=1}^{m}\left(y^{(i)} - \hat{p}\left(\vec{x}^{(i)T}\vec{\beta}^{[n-1]}\right)\right)x_1^{(i)} \\ \vdots \\ \sum_{i=1}^{m}\left(y^{(i)} - \hat{p}\left(\vec{x}^{(i)T}\vec{\beta}^{[n-1]}\right)\right)x_q^{(i)} \end{bmatrix} = X\left(\vec{y} - \vec{p}\left(X, \vec{\beta}^{[n-1]}\right)\right)$$

Where $\vec{p}\left(X,\vec{\beta}^{[n-1]}\right) = \begin{bmatrix} \hat{p}\left(\vec{x}^{\,(1)^T}\vec{\beta}^{[n-1]}\right) \\ \hat{p}\left(\vec{x}^{\,(2)^T}\vec{\beta}^{[n-1]}\right) \\ \vdots \\ \hat{p}\left(\vec{x}^{\,(m)^T}\vec{\beta}^{[n-1]}\right) \end{bmatrix}$

$\dfrac{\partial\ell^2\left(\vec{\beta}^{[n-1]}|\vec{x}\right)}{\partial{\vec{\beta}^{[n-1]}}^2} = \begin{bmatrix} \dfrac{\partial\ell^2\left(\vec{\beta}|\vec{x}\right)}{\partial\beta_0\partial\beta_0} & \dfrac{\partial\ell^2\left(\vec{\beta}|\vec{x}\right)}{\partial\beta_0\partial\beta_1} & \cdots & \dfrac{\partial\ell^2\left(\vec{\beta}|\vec{x}\right)}{\partial\beta_0\partial\beta_q} \\ \dfrac{\partial\ell^2\left(\vec{\beta}|\vec{x}\right)}{\partial\beta_1\partial\beta_0} & \dfrac{\partial\ell^2\left(\vec{\beta}|\vec{x}\right)}{\partial\beta_1\partial\beta_1} & \vdots & \dfrac{\partial\ell^2\left(\vec{\beta}|\vec{x}\right)}{\partial\beta_1\partial\beta_q} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial\ell^2\left(\vec{\beta}|\vec{x}\right)}{\partial\beta_q\partial\beta_0} & \dfrac{\partial\ell^2\left(\vec{\beta}|\vec{x}\right)}{\partial\beta_q\partial\beta_1} & \cdots & \dfrac{\partial\ell^2\left(\vec{\beta}|\vec{x}\right)}{\partial\beta_q\partial\beta_q} \end{bmatrix}$, where

$$\frac{\partial\ell^2\left(\vec{\beta}|\vec{x}\right)}{\partial\beta_j\partial\beta_k} = \frac{\partial}{\partial\beta_k}\left(\frac{\partial\ell\left(\vec{\beta}|\vec{x}\right)}{\partial\beta_j}\right) = \frac{\partial}{\partial\beta_k}\left(\sum_{i=1}^{m}\left(y^{(i)} - \frac{1}{1+e^{-\vec{x}^{(i)^T}\vec{\beta}}}\right)x_j^{(i)}\right)$$

$$= \sum_{i=1}^{m}\left(-\frac{-1}{\left(1+e^{-\vec{x}^{(i)^T}\vec{\beta}}\right)^2} \times -e^{-\vec{x}^{(i)^T}\vec{\beta}}\right)x_j^{(i)}x_k^{(i)} = -\sum_{i=1}^{m}\left(\frac{1}{1+e^{-\vec{x}^{(i)^T}\vec{\beta}}} \times \frac{e^{-\vec{x}^{(i)^T}\vec{\beta}}}{1+e^{-\vec{x}^{(i)^T}\vec{\beta}}}\right)x_j^{(i)}x_k^{(i)}$$

$$= -\sum_{i=1}^{m}\left(\frac{1}{1+e^{-\vec{x}^{(i)^T}\vec{\beta}}} \times \left(1 - \frac{1}{1+e^{-\vec{x}^{(i)^T}\vec{\beta}}}\right)\right)x_j^{(i)}x_k^{(i)}$$

$$= -\sum_{i=1}^{m}\hat{p}\left(\vec{x}^{\,(1)^T}\vec{\beta}^{[n-1]}\right) \times \left(1 - \hat{p}\left(\vec{x}^{\,(1)^T}\vec{\beta}^{[n-1]}\right)\right) \times x_j^{(i)} \times x_k^{(i)}$$

In matrix form,

$$\frac{\partial\ell^2\left(\vec{\beta}^{[n-1]}|\vec{x}\right)}{\partial{\vec{\beta}^{[n-1]}}^2} = -\sum_{i=1}^{m}\hat{p}\left(\vec{x}^{\,(1)^T}\vec{\beta}^{[n-1]}\right) \times \left(1 - \hat{p}\left(\vec{x}^{\,(1)^T}\vec{\beta}^{[n-1]}\right)\right) \times \vec{x}^{(i)} \times \vec{x}^{(i)^T}$$