Logistic Regression

Logistic function:

$$p(t) = \frac{1}{1+\exp(-t)} = \frac{\exp(t)}{1+\exp(t)}$$

It has the property $1-p(t) = p(-t)$.

We consider a discrete set of outcomes $y_i \in \{0,1\}$. We define the probability that $y_i = 1$ as

$$p(y_i = 1|x_i,\boldsymbol{\beta}) \equiv \frac{\exp\{\beta_0+\beta_1 x_i\}}{1+\exp\{\beta_0+\beta_1 x_i\}} \rightarrow p(y_i = 0|x_i,\boldsymbol{\beta}) = 1-p(y_i = 1|x_i,\boldsymbol{\beta})$$

The probability distribution for y_i may be written as

$$p(y_i|x_i,\boldsymbol{\beta}) = [p(y_i = 1|x_i,\boldsymbol{\beta})]^{y_i}[1-p(y_i = 1|x_i,\boldsymbol{\beta})]^{1-y_i}$$

The probability that we have a data set $\mathcal{D} = \{(x_i,y_i)\}$ is then

$$p(\mathcal{D}|\boldsymbol{\beta}) = \prod_{i=1}^{n}[p(y_i = 1|x_i,\boldsymbol{\beta})]^{y_i}[1-p(y_i = 1|x_i,\boldsymbol{\beta})]^{1-y_i}$$

We now select the vector of parameters $\boldsymbol{\beta}$ such that the probability for having the data set $\mathcal{D}$ is maximized. Alternatively, we can maximize the natural log of $p(\mathcal{D}|\boldsymbol{\beta})$

$$C(\boldsymbol{\beta}) \equiv \ln\{p(\mathcal{D}|\boldsymbol{\beta})\} = \sum_{i=1}^{n} y_i \ln\{p(y_i = 1|x_i,\boldsymbol{\beta})\} + (1-y_i)\ln\{1-p(y_i = 1|x_i,\boldsymbol{\beta})\}$$

which we define as the cost function. Since $1-p(t) = p(-t)$ we can write the cost function as (with $t = \beta_0+\beta_1 x_i$ as shorthand notation)

$$C(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \ln\left\{\frac{1}{1+\exp(-t)}\right\} + (1-y_i)\ln\left\{\frac{1}{1+\exp(t)}\right\}$$

$$= \sum_{i=1}^{n} -y_i \ln\{1+\exp(-t)\} + (y_i - 1)\ln\{1+\exp(t)\} = \sum_{i=1}^{n} y_i \ln\left\{\frac{1+\exp\{t\}}{1+\exp\{-t\}}\right\} - \ln\{1+\exp\{t\}\}$$

$$= \sum_{i=1}^{n} y_i t - \ln\{1+\exp\{t\}\} = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) - \ln\{1+\exp\{\beta_0+\beta_1 x_i\}\}$$

Alternatively, we may redefine the cost function as the negative natural log of the probability distribution of the data set $\mathcal{D}$

$$C(\boldsymbol{\beta}) \equiv -\sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) - \ln\{1+\exp\{\beta_0+\beta_1 x_i\}\}$$

and then find the minimum. This quantity is known in statistics as the **cross entropy**. The components of the gradient $\partial_{\boldsymbol{\beta}} C(\boldsymbol{\beta})$ are

$$\frac{\partial C(\boldsymbol{\beta})}{\partial \beta_0} = -\sum_{i=1}^{n} y_i - \frac{\exp\{\beta_0+\beta_1 x_i\}}{1+\exp\{\beta_0+\beta_1 x_i\}} = -\sum_{i=1}^{n} y_i - p(y_i = 1|x_i, \boldsymbol{\beta})$$

$$\frac{\partial C(\boldsymbol{\beta})}{\partial \beta_1} = -\sum_{i=1}^{n} x_i y_i - x_i\frac{\exp\{\beta_0+\beta_1 x_i\}}{1+\exp\{\beta_0+\beta_1 x_i\}} = -\sum_{i=1}^{n} x_i y_i - x_i p(y_i = 1|x_i, \boldsymbol{\beta})$$

If we define a vector $\mathbf{y}$ of n elements y_i, a vector $\mathbf{p}$ of n elements $p(y_i = 1|x_i, \boldsymbol{\beta})$ and the matrix

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

we can write the gradient of the cost function as

$$\frac{\partial C(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\mathbf{X}^T(\mathbf{y} - \mathbf{p})$$

The components of the Hessian matrix of $C(\boldsymbol{\beta})$ are

$$H_{ij} = \frac{\partial C(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_j}$$

$$= -\frac{\partial}{\partial \beta_i} \sum_k X_{jk}^T (y_k - p_k) = \sum_k X_{jk}^T \frac{\partial p_k}{\partial \beta_i}$$

where

$$\frac{\partial p_k}{\partial \beta_i} = \frac{\partial}{\partial \beta_i} \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 x_k)\}}$$

$$= \frac{\partial}{\partial \beta_i} \frac{1}{1 + \exp\left\{-\sum_l X_{kl}\beta_l\right\}} = -\frac{1}{\left(1 + \exp\left\{-\sum_l X_{kl}\beta_l\right\}\right)^2} \exp\left\{-\sum_l X_{kl}\beta_l\right\}(-X_{ki})$$

$$= \frac{\exp\{-(\beta_0 + \beta_1 x_k)\}}{(1 + \exp\{-(\beta_0 + \beta_1 x_k)\})^2} X_{ki}$$

Since
$p(t)[1 - p(t)] = p(t) - p(t)^2 = \dfrac{1}{1 + \exp(-t)} - \dfrac{1}{[1 + \exp(-t)]^2} = \dfrac{1 + \exp(-t) - 1}{[1 + \exp(-t)]^2} = \dfrac{\exp(-t)}{[1 + \exp(-t)]^2}$
we can write the components of the Hessian as

$$H_{ij} = \sum_k X_{jk}^T p_k (1 - p_k) X_{ki}$$

If $\mathbf{W}$ is a diagonal matrix we can write the Hessian on the form

$$\mathbf{H} = \mathbf{X}^T \mathbf{W} \mathbf{X}$$

since

$$H_{ij} = \sum_k X_{jk}^T (\mathbf{W}\mathbf{X})_{ki}$$

$$= \sum_k X^T_{jk} \sum_l W_{kl} \delta_{kl} X_{li} = \sum_k X^T_{jk} W_{kk} X_{ki}$$

Then we can identify $W_{ij} = p_i(1-p_i)\delta_{ij}$.

To set the gradient of the cost function $C(\boldsymbol{\beta})$ to zero, we can use Newton-Raphson's method

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \mathbf{J}[\mathbf{f}(\mathbf{x}_n)]^{-1}\mathbf{f}(\mathbf{x}_n)$$

where we substitute $\mathbf{x} \to \boldsymbol{\beta}$ and $\mathbf{f}(\mathbf{x}) \to \partial C(\boldsymbol{\beta}) \,/\, \partial \boldsymbol{\beta}$. This means that we also substitute

$$J_{ij} = \frac{\partial f_i}{\partial x_j} \to \frac{\partial}{\partial \beta_j} \frac{\partial}{\partial \beta_i} C(\boldsymbol{\beta}) = \frac{\partial^2 C(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_i} = H_{ij}$$

So in total, we end up with

$$\boldsymbol{\beta}_{n+1} = \boldsymbol{\beta}_n - \mathbf{H}[C(\boldsymbol{\beta}_n)]^{-1} \frac{\partial C(\boldsymbol{\beta}_n)}{\partial \boldsymbol{\beta}}$$