Logistic Regression

We consider $K = 2$ and encode the y-variable as 0 or 1. The *logistic regression model* is given by

$$\Pr(Y = 1 \mid X = x) = \frac{\exp((1, x^T)\beta)}{1 + \exp((1, x^T)\beta)}$$

Hence

$$\Pr(Y = 0 \mid X = x) = 1 - \frac{\exp((1, x^T)\beta)}{1 + \exp((1, x^T)\beta)} = \frac{1}{1 + \exp((1, x^T)\beta)}.$$

We saw that the conditional distribution of Y given X in the LDA setup is a logistic regression model.

Logistic Regression – Notation

Given a dataset $(y_1, x_1), \ldots, (y_N, x_N)$ write

$$\mathbf{p}(\beta) = (p_i(\beta))_{i=1}^N, \quad p_i(\beta) = \frac{\exp((1, x_i^T)\beta)}{1 + \exp((1, x_i^T)\beta)}.$$

With $h : \mathbb{R}^N \to \mathbb{R}^N$

$$h_i(z) = -\log(1 + \exp(z_i))$$

and taking coordinatewise logarithm

$$\log \mathbf{p}(\beta) = \mathbf{X}\beta + h(\mathbf{X}\beta)$$

and

$$\log(\mathbf{1} - \mathbf{p}(\beta)) = h(\mathbf{X}\beta)$$

Logistic Regression – The Minus-Log-Likelihood Function

The (conditional) likelihood function of observing $y_1, \ldots, y_N$ given $x_1, \ldots, x_N$ is

$$\mathcal{L}(\beta) = \prod_{i=1}^N p_i(\beta)^{y_i}(1 - p_i(\beta))^{1-y_i}$$

and the minus-log-likelihood function is

$$\begin{aligned} l(\beta) &= -\mathbf{y}^T(\mathbf{X}\beta + h(\mathbf{X}\beta)) - (\mathbf{1} - \mathbf{y})^T h(\mathbf{X}\beta) \\ &= -\mathbf{y}^T\mathbf{X}\beta - \mathbf{1}^T h(\mathbf{X}\beta) \end{aligned}$$

Observe that $D_z h(z)$ is diagonal with

$$D_z h(z)_{ii} = -\frac{\exp(z_i)}{1 + \exp(z_i)}$$

Logistic Regression – The MLE

By differentiation

$$\begin{aligned} D_\beta l(\beta) &= -\mathbf{y}^T\mathbf{X} - \mathbf{1}^T D_z h(\mathbf{X}\beta)\mathbf{X} \\ &= -\mathbf{y}^T\mathbf{X} + \mathbf{p}(\beta)^T\mathbf{X} \\ &= (\mathbf{p}(\beta)^T - \mathbf{y}^T)\mathbf{X} \end{aligned}$$

and

$$D_\beta^2 l(\beta) = D_\beta \mathbf{p}(\beta)^T\mathbf{X} = \mathbf{X}^T\mathbf{W}(\beta)\mathbf{X}$$

with

$$\begin{aligned} \mathbf{W}(\beta) &= \mathrm{diag}(\mathbf{p}(\beta))\mathrm{diag}(1-\mathbf{p}(\beta)) \\ &= \left\{ \begin{array}{ccc} p_1(\beta)(1-p_1(\beta)) & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & p_N(\beta)(1-p_N(\beta)) \end{array} \right\} \end{aligned}$$

Likelihood Equation

The non-linear likelihood estimation equation reads (after transposition)

$$\mathbf{X}^T\mathbf{p}(\beta) = \mathbf{X}^T\mathbf{y}$$

Since $D_\beta^2 l(\beta) = \mathbf{X}^T\mathbf{W}(\beta)\mathbf{X}$ is *positive definite* whenever $\mathbf{X}$ has full rank $p+1$, the minus-log-likelihood function is strictly convex and a minimum is unique.

There is *no solution* if the x-values for the two groups can be separated completely by a hyperplane.

Logistic Regression – Algorithm

A first order Taylor expansion

$$\mathbf{p}(\beta) \simeq \mathbf{p}(\beta_0) + \mathbf{W}(\beta_0)\mathbf{X}(\beta - \beta_0)$$

around β_0 yields the approximating equation

$$\mathbf{X}^T\mathbf{W}(\beta_0)\mathbf{X}\beta = \mathbf{X}^T\mathbf{W}(\beta_0)(\underbrace{\mathbf{X}\beta_0 + \mathbf{W}(\beta_0)^{-1}(\mathbf{y} - \mathbf{p}(\beta_0))}_{\text{adjusted response}=\mathbf{z}_0}).$$

The solution is precisely the solutions of the *weighted least squares problem*

$$\operatorname*{argmin}_\beta\ (\mathbf{z}_0 - \mathbf{X}\beta)^T\mathbf{W}(\beta_0)(\mathbf{z}_0 - \mathbf{X}\beta)$$

Iteration yielding a sequence β_n, $n \geq 0$, is known as the *iterative reweighted least squares* algorithm – or IRLS – using the *adjusted response*

$$\mathbf{z}_n = \mathbf{X}\beta_n + \mathbf{W}(\beta_n)^{-1}(\mathbf{y} - \mathbf{p}(\beta_n))$$

in the $(n+1)$'th iteration. The algorithm is equivalent to the Newton-Raphson algorithm.

Figure 4.12 – South African Heart Disease Data

A typical use of *logistic regression.* The response variable is Myocardial Infarction. The two cases (0/1) are color coded in the plot.

The plot reveals pair-wise – and marginal – effects of the 7 observed variables on MI.

And clear correlations between `obesity` and `sbp` (systolic blood pressure), say.

Multinomial Regression and LDA

It is possible to formulate a multinomial version of the binary logistic regression model.

The algorithm for estimation becomes more complicated.

LDA implements the *plug-in principle* using MLE for the full parameter. Logistic/multinomial regression implements the *conditional plug-in principle* using MLE in the conditional distribution.

Logistic regression makes fewer distributional assumptions. Deviations from normality *could* affect LDA in the negative direction.

If the distributional assumptions of LDA are fulfilled LDA is a little more efficient.