Maximum Likelihood in Logistic Regression - Logit Algorithm

Recall : Log-likelihood

$$\ell(\beta) = \sum_i y_i \cdot \log\left(F(\beta^T x_i)\right) + (1-y_i)\log\left(1-F(\beta^T x_i)\right)$$

where $F(u) = \frac{1}{1+e^{-u}}$

$$F'(u) = F(u)\cdot(1-F(u))$$

$\Rightarrow$ gradient and Hessian matrix

$$D_\ell(\beta) = \sum_i \left\{ \left(y_i - F(\beta^T x_i)\right)\cdot \underbrace{x_i}_{\text{vector}} \right\}$$

$$H_\ell(\beta) = -\sum_i \left\{ \underbrace{F(\beta^T x_i)\cdot(1-F(\beta^T x_i))}_{w_i \text{ "weights"}} \cdot \underbrace{x_i x_i^T}_{\text{matrix}} \right\}$$

design matrix $\mathcal{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}$

weight matrix $W = \begin{pmatrix} w_1 & & 0 \\ & \ddots & \\ 0 & & w_n \end{pmatrix}$

$\Rightarrow$ we had seen:

$$H_\ell(\beta) = -\mathcal{X}^T W \mathcal{X}$$

$\Rightarrow$ do this similarly for the gradient : $D_\ell(\beta) = \mathcal{X}^T W \tilde{y}$ (See slide 69)

$\Rightarrow$ we had a nonlinear system of equations:

$D_\ell(\beta) \overset{!}{=} 0$ to solve

$\Rightarrow$ use Newton Raphson: (slide 66)

$$\hat{\beta}^{new} = \hat{\beta}^{old} - \mathcal{H}_\ell(\hat{\beta}^{old})^{-1} \cdot D_\ell(\hat{\beta}^{old})$$

$$= \hat{\beta}^{old} - (-X^T W X)^{-1} X^T W \tilde{y}$$

$$= \hat{\beta}^{old} + (X^T W X)^{-1} X^T W \tilde{y}$$

contain y_i and $\hat{\beta}^{old}$

$$= \ldots \text{ (see slide 70)}$$

in each iteration step:

$$\hat{\beta}^{new} = (X^T W X)^{-1} X^T W z$$

adjusted dependent variable

$\Rightarrow$ weighted least squares(*) estimation in each iteration step

$\hat{=}$ $\boxed{\text{IRLS algorithm}}$

iteratively reweighted least squares

(*) for comparison least squares in linear regression

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Algorithm:

- initial values are needed for z (not β) (see slide 71)

$$z = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix} = \underbrace{X \hat{\beta}^{old}}_{=: \hat{\eta}_i^{old}} + \tilde{y}$$

we could use here:

initial values for μ_i $= F(\eta_i)$ $\longrightarrow$

$$\hat{\mu}_i^{(0)} = \frac{y_i}{2} + \frac{1}{4} = \begin{cases} 1/4 & y_i = 0 \\ 3/4 & y_i = 1 \end{cases}$$

$$\hat{\eta}_i^{(0)} = F^{-1}(\hat{\mu}_i^{(0)})$$

- Stop the iteration if:

(1) relative change of the coefficients is small

$$\frac{\| \hat{\beta}^{new} - \hat{\beta}^{old} \|}{\| \hat{\beta}^{old} \|} < \varepsilon \quad \text{tolerance}$$

or:

(2) relative change of objective criterion (here: log likelihood) is small

$$\frac{| \ell(\hat{\beta}^{new}) - \ell(\hat{\beta}^{old}) |}{| \ell(\hat{\beta}^{old}) |} < \varepsilon$$

What does "logit" mean?

chance or odds: $\dfrac{p}{1-p}$

$$\Longrightarrow \operatorname{logit}(p) = \underset{\ln}{\log}\left(\frac{p}{1-p}\right) = F^{-1}(p)$$

inverse function: $u = \log\left(\dfrac{p}{1-p}\right)$

$$\Longleftrightarrow e^u = \frac{p}{1-p} \Longleftrightarrow e^u(1-p) = p$$

$$\Longleftrightarrow e^u = p(1+e^u) \Longleftrightarrow p = \frac{e^u}{1+e^u}$$

$$\Longrightarrow F(u) = \frac{e^u}{1+e^u} = \frac{1}{1+e^{-u}} \;\hat{=}\; \text{inverse of logit transformation}$$