Derivation of the Maximum Likelihood Estimator in Logistic Regression

Assume : $Y_1, \dots, Y_n \underset{iid}{\sim} B(1, p)$

independently and identically distributed

$\Longrightarrow$ likelihood function

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

product as we have independence

$\Longrightarrow$ log likelihood function

$$\ell(p) = \log(L(p))$$

in statistics and in R: $\log = \ln$

$$= \sum_{i=1}^{n} y_i \cdot \log(p) + (1-y_i) \cdot \log(1-p)$$

$\Longrightarrow$ derivative

$$\ell'(p) = \sum_i \frac{y_i}{p} - \frac{1-y_i}{1-p}$$

$$\Rightarrow \ell'(p) = \sum_i \frac{y_i - \cancel{y_i p} - p + \cancel{y_i p}}{p(1-p)}$$

$$= \sum_i \frac{y_i - p}{p(1-p)}$$

Using explanatory variables x_i :

$$Y_i \sim B(1, p_{x_i})$$

where $p_{x_i} = F(\beta^T x_i)$

and $F(u) = \dfrac{1}{1+e^{-u}}$

it can be shown :

$$F'(u) = f(u) \qquad \left(\begin{array}{l} F \mathrel{\hat{=}} \text{cdf} \\ \Rightarrow f \mathrel{\hat{=}} \text{pdf} \end{array} \right)$$

$$F'(u) = \frac{e^{-u}}{(1+e^{-u})^2}$$

$$\Longrightarrow F'(u) = \underbrace{\frac{1}{1+e^{-u}}}_{=F(u)} \cdot \underbrace{\frac{e^{-u}}{1+e^{-u}}}_{=1-F(u)}$$

so also: $F'(u) = F(u)\cdot(1-F(u))$

The log likelihood becomes:

(replace p in $\ell(p)$ by p_{x_i})

$$\ell(\beta) = \sum_i y_i \log(p_{x_i}) + (1-y_i)\log(1-p_{x_i})$$

we aim to find/estimate β

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

in the multidimensional case 1st derivative $\hat{=}$ gradient

$$D_\ell(\beta) = \frac{\partial \ell}{\partial \beta} = \sum_i \frac{y_i - p_{x_i}}{p_{x_i}(1-p_{x_i})} \cdot \frac{\partial p_{x_i}}{\partial \beta}$$

$$= \sum_i \underbrace{\frac{y_i - F(\beta^T x_i)}{F(\beta^T x_i)\cdot(1-F(\beta^T x_i))}}_{\in \mathbb{R}} \cdot \underbrace{F'(\beta^T x_i)}_{\in \mathbb{R}} \cdot x_i$$

$(p+1)\times 1$ vector

remark: up to here everything also holds for probit, i.e. $F = \Phi$

for logit we have additionally: $F'(u) = F(u)\cdot(1-F(u))$

$$\Rightarrow \quad \underbrace{D_\ell(\beta)}_{\text{vector}} = \sum_i \Big\{ \underbrace{\big(y_i - F(\beta^T x_i)\big)}_{\in \mathbb{R}} \cdot \underset{\text{vector}}{x_i} \Big\}$$

Now as in the linear model:

$$D_\ell(\beta) \overset{!}{=} 0 \quad (\text{solve for } \beta)$$

$\Rightarrow$ determine the root of $D_\ell(\beta)$

$\Rightarrow$ Newton's method

+ check if "2nd derivative" $\hat{=}$ Hessian is negative definite $\Rightarrow$ maximum!

Slide 67:

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

$\Rightarrow$ derive the Hessian matrix:

$$D_\ell(\beta) = \sum_i \big\{ (y_i - F(\beta^T x_i)) \cdot x_i \big\}$$

$$= \frac{\partial \ell}{\partial \beta}$$

$$\mathcal{H}_\ell(\beta) = \frac{\partial^2 \ell}{\partial \beta \, \partial \beta^T}$$

Hessian

$$= \sum_i \big\{ -F'(\beta^T x_i) \cdot x_i \cdot x_i^T \big\}$$

$$= - \sum_{i=1}^{n} \Big\{ \underbrace{F'(\beta^T x_i)}_{F(\beta^T x_i)\cdot(1-F(\beta^T x_i)) = w_i} \cdot x_i x_i^T \Big\}$$

weights > 0

recall from linear model:

$$x_i = \begin{pmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix}$$

$$\Rightarrow \underset{(p+1)\times 1}{x_i}\,\underset{1\times(p+1)}{x_i^T} = \underbrace{\begin{pmatrix} 1 & x_{i1} & \cdots & x_{ip} \\ x_{i1} & x_{i1}^2 & \cdots & x_{i1}x_{ip} \\ \vdots & & \ddots & \vdots \\ x_{ip} & x_{i1}x_{ip} & \cdots & x_{ip}^2 \end{pmatrix}}_{(p+1)\times(p+1)}$$

$$\Rightarrow \sum_i x_i x_i^T = \mathcal{X}^T \mathcal{X}$$

<u>now:</u> $\sum_i w_i x_i x_i^T = \mathcal{X}^T W \mathcal{X}$

we know from the linear model: $\mathcal{X}^T\mathcal{X} > 0$ positive definite

$\Rightarrow$ as we have $w_i > 0$ then also:

$$\mathcal{X}^T W \mathcal{X} > 0 \leftarrow \text{positive definite}$$

$\Rightarrow$ so for our Hessian:

minus!

$$H_\ell(\beta) = -\sum_i w_i x_i x_i^T$$

$$= -\mathcal{X}^T W \mathcal{X} < 0$$

negative definite

$\Rightarrow$ <u>maximum!</u>

$\Rightarrow$ our estimate $\hat{\beta}$ found by the Newton-Raphson method (slide 67) is <u>the maximum likelihood estimate!</u>