Logistic regression

EKT 720

Introduction to Statistical learning

September 2019

1 Logistic regression

The logistic regression model with binary response,

$$\begin{aligned}
\Pr\left(y_i = 1|\ x_i, \boldsymbol{\beta}\right) &= \frac{e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}} \\
&= \frac{1}{1+e^{-\boldsymbol{x}_i^T\boldsymbol{\beta}}} \\
&= p\left(\boldsymbol{x}_i;\boldsymbol{\beta}\right) \qquad (1)
\end{aligned}$$

with $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix}$ a $n \times p$ matrix, $\boldsymbol{x}_i^T$, a $1 \times p$ vector and $\boldsymbol{\beta}$ a $p \times 1$ vector of parameters, for $i = 1, 2, \ldots, n$.

Equation 1 can be linearised as follows,

$$\begin{aligned}
\Pr\left(y_i = 1|\boldsymbol{x}_i, \boldsymbol{\beta}\right) &= \frac{1}{1+e^{-\boldsymbol{x}_i^T\boldsymbol{\beta}}} \\
odds_i &= \frac{p\left(\boldsymbol{x}_i;\boldsymbol{\beta}\right)}{1-p\left(\boldsymbol{x}_i;\boldsymbol{\beta}\right)} \\
&= \frac{\frac{1}{1+e^{-\boldsymbol{x}_i^T\boldsymbol{\beta}}}}{1-\frac{1}{1+e^{-\boldsymbol{x}_i^T\boldsymbol{\beta}}}} \\
&= e^{\boldsymbol{x}_i^T\boldsymbol{\beta}} \qquad (2) \\
log(odds_i) &= \boldsymbol{x}_i^T\boldsymbol{\beta} \qquad (3)
\end{aligned}$$

2 Log-likelihood function

Consider a random sample of size n, $\left(\boldsymbol{x}_i^T, y_i\right)$, for $i = 1, 2, \ldots n$, then the likelihood function of $\boldsymbol{\beta}$ under the assumption of independence is,

$$L(\boldsymbol{\beta}|\boldsymbol{X}) = \prod_{i=1}^{n} p\left(\boldsymbol{x}_i;\boldsymbol{\beta}\right)$$

with the log likelihood function

$$
\begin{aligned}
l(\boldsymbol{\beta}|\boldsymbol{X}) &= \sum\nolimits_{i=1}^{n} \log p(\boldsymbol{x}_i;\boldsymbol{\beta}) \\
&= \sum\nolimits_{i=1}^{n} \{y_i \log(p(\boldsymbol{x}_i;\boldsymbol{\beta})) + (1-y_i)\log(1-p(\boldsymbol{x}_i;\boldsymbol{\beta}))\} \\
&= \sum\nolimits_{i=1}^{n} \{y_i \log(p(\boldsymbol{x}_i;\boldsymbol{\beta})) + \log(1-p(\boldsymbol{x}_i;\boldsymbol{\beta})) - y_i\log(1-p(\boldsymbol{x}_i;\boldsymbol{\beta}))\} \\
&= \sum\nolimits_{i=1}^{n} \left\{y_i \log\left(\frac{p(\boldsymbol{x}_i;\boldsymbol{\beta})}{(1-p(\boldsymbol{x}_i;\boldsymbol{\beta}))}\right) + \log(1-p(\boldsymbol{x}_i;\boldsymbol{\beta}))\right\} \\
&= \sum\nolimits_{i=1}^{n} \left\{y_i \boldsymbol{x}_i^T\boldsymbol{\beta} + \log(1-p(\boldsymbol{x}_i;\boldsymbol{\beta}))\right\} \\
&= \sum\nolimits_{i=1}^{n} \left\{y_i \boldsymbol{x}_i^T\boldsymbol{\beta} + \log\left(1-\frac{e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}\right)\right\} \\
&= \sum\nolimits_{i=1}^{n} \left\{y_i \boldsymbol{x}_i^T\boldsymbol{\beta} + \log\left(\frac{1}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}\right)\right\} \\
&= \sum\nolimits_{i=1}^{n} \left\{y_i \boldsymbol{x}_i^T\boldsymbol{\beta} - \log\left(1+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\right)\right\} \qquad (4)
\end{aligned}
$$

3 Maximum likelihood estimation

We maximise Equation 4 using the Newton Raphson algorithm. This requires the first derivatives, the score or gradient function

$$
\begin{aligned}
\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \sum\nolimits_{i=1}^{n} \left\{y_i\boldsymbol{x}_i - \frac{e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}{\left(1+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\right)}\boldsymbol{x}_i\right\} \\
&= \sum\nolimits_{i=1}^{n} \{y_i\boldsymbol{x}_i - p(\boldsymbol{x}_i;\boldsymbol{\beta})\boldsymbol{x}_i\} \\
&= \sum\nolimits_{i=1}^{n} \{(y_i - p(\boldsymbol{x}_i;\boldsymbol{\beta}))\boldsymbol{x}_i\}
\end{aligned}
$$

or in matrix notation $\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{p})$, and the second derivatives or Hessian matrix

$$
\frac{\partial l^2(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} = -\sum\nolimits_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^T p(\boldsymbol{x}_i;\boldsymbol{\beta})(1-p(\boldsymbol{x}_i;\boldsymbol{\beta}))
$$

or in matrix notation $\frac{\partial l^2(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} = -\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$, with $\boldsymbol{W}$ a diagonal matrix with elements $p(\boldsymbol{x}_i;\boldsymbol{\beta})(1-p(\boldsymbol{x}_i;\boldsymbol{\beta}))$ as i^{th} diagonal element.

Estimating the parameters using Newton Raphson yields:

$$\begin{aligned}
\boldsymbol{\beta}^{new} &= \boldsymbol{\beta}^{old} + \left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\left(\boldsymbol{y}-\boldsymbol{p}\right) \\
&= \left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)\boldsymbol{\beta}^{old} + \left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\left(\boldsymbol{y}-\boldsymbol{p}\right) \\
&= \left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}\left(\boldsymbol{X}\boldsymbol{\beta}^{old} + \boldsymbol{W}^{-1}\left(\boldsymbol{y}-\boldsymbol{p}\right)\right) \\
&= \left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{z}
\end{aligned}$$

which is an $\boxed{\text{iteratively reweighted least squares (IRLS)}}$ solution to $\boldsymbol{\beta}$ with adjusted response $\boldsymbol{z} = \left(\boldsymbol{X}\boldsymbol{\beta}^{old} + \boldsymbol{W}^{-1}\left(\boldsymbol{y}-\boldsymbol{p}\right)\right)$.

4 IRLS Algorithm

The IRLS algorithm used to estimate the parameters $\boldsymbol{\beta}$ is given below

Algorithm 1 IRLS - binary logistic regression.

1. Select initial values for the regression parameters $\boldsymbol{\beta}^{old}$
2. Calculate the $p(\boldsymbol{x}_i, \boldsymbol{\beta}^{old}) = \frac{1}{1+e^{-\boldsymbol{x}_i^T\boldsymbol{\beta}^{old}}}$, $i = 1, \ldots, n$
3. Calculate the diagonal weight matrix W with elements $p(x_i, \boldsymbol{\beta}^{old})(1 - p(x_i, \boldsymbol{\beta}^{old}))$.
4. Calculate the Gradient vector and Hessian matrix
   (a) $\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^T\left(\boldsymbol{y}-\boldsymbol{p}\right)$
   (b) $\frac{\partial l^2(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} = -\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$
5. Calculate $\boldsymbol{\beta}^{new} = \boldsymbol{\beta}^{old} + (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{p})$ or
   $\boldsymbol{\beta}^{new} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{z}$, with adjusted response $\boldsymbol{z} = \left(\boldsymbol{X}\boldsymbol{\beta}^{old} + \boldsymbol{W}^{-1}\left(\boldsymbol{y}-\boldsymbol{p}\right)\right)$
6. Set $\boldsymbol{\beta}^{old} = \boldsymbol{\beta}^{new}$
7. Repeat steps (2) to (6) untill convergence.