Econometrics 720

1. Logistic Regression

The logistic regression model arises from the desire to model posterior probabilities of K classes via linear functions in x, while at the same time ensuring that they sum to one and remain in the interval $[0, 1]$.

A model that complies to the above is:

$$\begin{aligned}
log\frac{Pr(G=1|X=x)}{Pr(G=K|X=x)} &= \beta_{10} + \beta_1^T x \\
log(odds_1) &= \beta_{10} + \beta_1^T x \quad \text{where} \quad odds_1 = \frac{Pr(G=1|X=x)}{Pr(G=K|X=x)} \\
odds_1 &= e^{\beta_{10} + \beta_1^T x}
\end{aligned}$$

$$\begin{aligned}
log\frac{Pr(G=2|X=x)}{Pr(G=K|X=x)} &= \beta_{20} + \beta_2^T x \\
log(odds_2) &= \beta_{20} + \beta_2^T x \quad \text{where} \quad odds_2 = \frac{Pr(G=1|X=x)}{Pr(G=K|X=x)} \\
odds_2 &= e^{\beta_{20} + \beta_2^T x}
\end{aligned}$$

$\ldots$

$\ldots$

$$\begin{aligned}
log\frac{Pr(G=K-1|X=x)}{Pr(G=K|X=x)} &= \beta_{(K-1)0} + \beta_{K-1}^T x \\
log(odds_{K-1}) &= \beta_{(K-1)0} + \beta_{K-1}^T x \quad \text{where} \quad odds_{K-1} = \frac{Pr(G=K-1|X=x)}{Pr(G=K|X=x)} \\
odds_{K-1} &= e^{\beta_{(K-1)0} + \beta_{K-1}^T x}
\end{aligned}$$

The above imply the use of linear decision boundaries, based on the CDF of a logistic distribution.

Consider the following:

$$\begin{aligned}
\Sigma_{l=1}^{K-1} log(odds_l) &= \Sigma_{l=1}^{K-1} \beta_{l0} + \beta_l^T x \\
\frac{1}{Pr(G=K|X=x)} \Sigma_{l=1}^{K-1} Pr(G=l|X=x) &= \Sigma_{l=1}^{K-1} e^{\beta_{l0} + \beta_l^T x} \\
\frac{1}{Pr(G=K|X=x)} (1 - Pr(G=K|X=x)) &= \Sigma_{l=1}^{K-1} e^{\beta_{l0} + \beta_l^T x} \\
\frac{1}{Pr(G=K|X=x)} &= 1 + \Sigma_{l=1}^{K-1} e^{\beta_{l0} + \beta_l^T x} \\
Pr(G=K|X=x) &= \frac{1}{1 + \Sigma_{l=1}^{K-1} e^{\beta_{l0} + \beta_l^T x}}
\end{aligned}$$

$$\begin{aligned}
Pr(G=k|X=x) &= e^{\beta_{k0} + \beta_k^T x} Pr(G=K|X=x) \\
Pr(G=k|X=x) &= e^{\beta_{k0} + \beta_k^T x} \frac{1}{1 + \Sigma_{l=1}^{K-1} e^{\beta_{l0} + \beta_l^T x}}
\end{aligned}$$

Clearly $\Sigma_{l=1}^{K} Pr(G=l|X=x)$ is equal to 1, and all probabilities depend on the full parameter set $\theta = \{\beta_{10}, \beta_1, \beta_{20}, \beta_2, \ldots, \beta_{(K-1)0}, \beta_{K-1}\}$. These probabilities are denoted by $Pr(G=k|x=x) = p_k(x;\theta)$

Estimation of Logistic Regression models using Newton-Raphson

Objective: Estimate the parameters that maximize the conditional likelihood of G given X, using the modelling/training data.

The conditional log-likelihood function

- Denote $p_k(x_i;\theta) = Pr(G = k|X = X_i;\theta)$.
- Given the first input x_1, the posterior probability if its class being g_1 is $Pr(G = g_1|X = x_1)$.
- Assuming independence of observations, the posterior probability for the N observations each having class $g_i = 1, 2, \ldots, N$, given their inputs $x_1, x_2, \ldots, x_N$ is:

$$\Pi_{i=1}^{N} Pr(G = g_i|X = x_i)$$

  and therefore the conditional log likelihood is given as

$$\begin{aligned} l(\theta) &= \Sigma_{i=1}^{N} log Pr(G = g_i|X = x_i) \\ &= \Sigma_{i=1}^{N} log p_{g_i}(x_i;\theta) \end{aligned}$$

  where $p_{g_i}(x_i;\theta)$ is the probability of being in the group g that is associated with the i'th observation.
- Consider only cases where $K = 2$, binary classification. The log-likelihood function can be simplified by using the following:

$$\begin{aligned} y_i &= 1 && \text{when} & g_i &= 1 \\ y_i &= 0 && \text{when} & g_i &= 2 \\ p_1(x;\theta) &= p(x;\theta) \\ p_2(x;\theta) &= 1 - p(x;\theta) && \text{for two groups} \end{aligned}$$

- Since $K = 2$, the parameters $\theta = \beta_{10}, \beta_1$. Denote $\beta = (\beta_{10}, \beta_1)'$
- If $y_i = 1$, i.e, $g_i = 1$,

$$\begin{aligned} log p_{g_i}(x;\beta) &= log p_1(x;\beta) \\ &= 1.log p_1(x;\beta) \\ &= y_i log p_1(x;\beta) \end{aligned}$$

  If $y_i = 0$, i.e, $g_i = 0$,

$$\begin{aligned} log p_{g_i}(x;\beta) &= log p_2(x;\beta) \\ &= 1.log(1 - p_1(x;\beta)) \\ &= (1 - y_i) log(1 - p_1(x;\beta)) \end{aligned}$$

Since $y_i = 0$ or $(1 - y_i) = 0$ we have

$logp_{g_i}(x;\beta) = y_i logp_1(x;\beta) + (1 - y_i)log(1 - p_1(x;\beta))$

- The conditional log-likelihood function then is:

$$
\begin{aligned}
l(\beta) &= \Sigma_{i=1}^{N} \left[y_i \, logp(x_i,\beta) + (1-y_i) log(1 - p(x_i;\beta))\right] \\
&= \Sigma_{i=1}^{N} \left[y_i \, logp(x_i,\beta) + log(1 - p(x_i;\beta)) - y_i \, log(1 - p(x_i;\beta))\right] \\
&= \Sigma_{i=1}^{N} \left[y_i \, log\frac{p(x_i,\beta)}{(1 - p(x_i;\beta)} + log(1 - p(x_i;\beta)\right] \\
&= \Sigma_{i=1}^{N} \left[y_i \, \beta^T x_i + log(1 - \frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}})\right] \\
& \quad \text{with} \quad \beta = \{\beta_{10}, \beta_1\} \quad \text{and} \quad x \quad \text{coded to include the intercept} \\
&= \Sigma_{i=1}^{N} \left[y_i \, \beta^T x_i + log(\frac{1}{1 + e^{\beta^T x_i}})\right] \\
&= \Sigma_{i=1}^{N} \left[y_i \, \beta^T x_i - log(1 + e^{\beta^T x_i})\right]
\end{aligned}
$$

$$
\text{with } \beta = \begin{pmatrix} \beta_{10} \\ \beta_{11} \\ \beta_{12} \\ . \\ . \\ . \\ \beta_{1p} \end{pmatrix} \text{ and } x = \begin{pmatrix} 1 \\ x_{,1} \\ x_{,2} \\ . \\ . \\ . \\ x_{,p} \end{pmatrix}
$$

Maximum Likelihood Estimation

In order to maximise the log-likelihood function we set its derivatives equal to zero.

$$\frac{\partial l(\beta)}{\partial \beta} = \Sigma_{i=1}^{N} x_i (y_i - p(x_i;\beta)) = 0$$

which are $p + 1$ score equations *non-linear* in β.

In order to solve the score equations , we use the Newton-Raphson algorithm, which requires the second derivative or Hessian matrix

$$\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^t} = -\Sigma_{i=1}^{N} x_i x_i^T p(x_i;\beta)(1 - p(x_i;\beta))$$

The Newton-Raphson algorithm relates to the following update formula:

$$\beta^{new} = \beta^{old} - \left(\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^t}\right)^{-1} \frac{\partial l(\beta)}{\partial\beta}$$

where the derivatives are evaluated at β^{old}.

In matrix notation we get:

$$\begin{aligned}\frac{\partial l(\beta)}{\partial\beta} &= X^T(y-p)\\ \frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^t} &= -X^T\,W\,X\end{aligned}$$

with

y	the vector of y_i values
X	the $N \times (p+1)$ matrix of x_i values
p	the vector of fitted probabilities with i'th element $p(x_i;\beta^{old})$
W	a $N \times N$ diagonal matrix of weights with i'th diagonal element $p(x_i;\beta^{old})(1-p(x_i;\beta^{old}))$

The Newton-Raphson update formula is:

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

where

$$z = (X\beta^{old} + W^{-1}(y-p))$$

The vector z is sometimes called the adjusted response, and the equations are solved repeatedly.