Logistic regression

Department of Statistics, University of South Carolina

Stat 705: Data Analysis II

In OLS, we have

$$\operatorname{argmin}_{\beta} \sum_i (y_i - x_i \boldsymbol{\beta})^2,$$

$$\frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{\beta}} = -2 \sum_i (y_i - x_i \boldsymbol{\beta}) x_i = 0$$

This is a linear system with p equations and p unknowns. So it can be solved using standard linear algebra theory with a closed form solution.

Likelihood

The logistic regression model can be written as

$$\log \frac{p}{1-p} = \mathbf{X}\boldsymbol{\beta}$$

Hence,

$$p = \frac{e^{\mathbf{X}\beta}}{1+e^{\mathbf{X}\boldsymbol{\beta}}}$$

The likelihood function for logistic regression is

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

The Score Function of Logistic Regression

$$\begin{aligned}
\log L(\beta) &= \ell(\beta) = \sum_i^n [y_i \log p_i + (1 - y_i) \log(1 - p_i)] \\
&= \sum_i^n [y_i \beta^T X_i - \log(1 + e^{\beta^T X_i})] \\
\frac{\partial \ell}{\partial \beta} &= \sum_i X_i (y_i - p_i) = 0
\end{aligned}$$

In matrix form can be expressed as:

$$\begin{aligned}
\frac{\partial \ell}{\partial \beta} &= X^T (y - p) \qquad \text{Score Function} \\
\frac{\partial^2 \ell}{\partial^2 \beta} &= -X^T W X,
\end{aligned}$$

where $W = \text{diag}[p_i(1 - p_i)]$.

How to get the estimates?

Newton-Raphson in one dimension: Say we want to find where $f(x) = 0$ for differentiable $f(x)$. Let x_0 be such that $f(x_0) = 0$. Taylor's theorem tells us

$$f(x_0) \approx f(x) + f'(x)(x_0 - x).$$

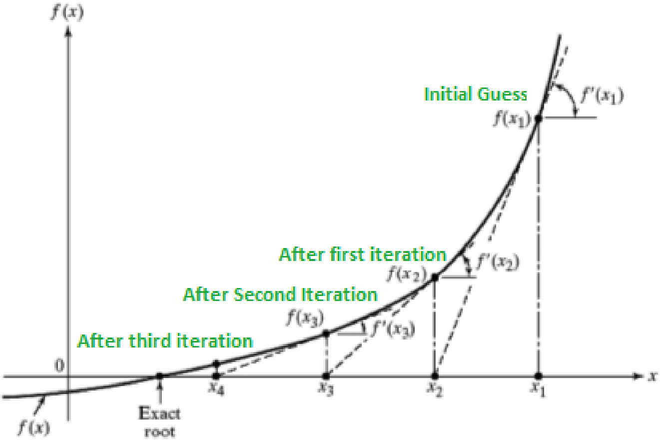
Plugging in $f(x_0) = 0$ and solving for x_0 we get $\hat{x}_0 = x - \frac{f(x)}{f'(x)}$. Starting at an x near x_0, $\hat{x}_0$ should be closer to x_0 than x was. Let's iterate this idea t times:

$$x^{(t+1)} = x^{(t)} - \frac{f(x^{(t)})}{f'(x^{(t)})}.$$

Eventually, if things go right, $x^{(t)}$ should be close to x_0.

Newton-Raphson

$$x^{(t+1)} = x^{(t)} - \frac{f(x^{(t)})}{f'(x^{(t)})}.$$



If $\mathbf{f}(\mathbf{x}) : \mathbb{R}^p \to \mathbb{R}^p$, the idea works the same, but in vector/matrix terms. Start with an initial guess $\mathbf{x}^{(0)}$ and iterate

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - [D\mathbf{f}(\mathbf{x}^{(t)})]^{-1}\mathbf{f}(\mathbf{x}^{(t)}).$$

If things are "done right," then this should converge to $\mathbf{x}_0$ such that $\mathbf{f}(\mathbf{x}_0) = \mathbf{0}$.
We are interested in solving $DL(\beta) = \mathbf{0}$ (the score, or likelihood equations!) where

$$DL(\beta) = \begin{bmatrix} \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_1} \\ \vdots \\ \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_p} \end{bmatrix} \text{ and } D^2L(\beta) = \begin{bmatrix} \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_1^2} & \cdots & \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_1 \partial \beta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_p \partial \beta_1} & \cdots & \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_p^2} \end{bmatrix}.$$

Newton-Raphson

So for us, we start with $\boldsymbol{\beta}^{(0)}$ (maybe through a MOM or least squares estimate) and iterate

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - [D^2 L(\boldsymbol{\beta})(\boldsymbol{\beta}^{(t)})]^{-1} DL(\boldsymbol{\beta}^{(t)}).$$

The process is typically stopped when $|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}| < \epsilon$.

- Newton-Raphson uses $D^2 L(\boldsymbol{\beta})$ as is, with the **y** plugged in.
- Fisher scoring instead uses $E\{D^2 L(\boldsymbol{\beta})\}$, with expectation taken over **Y**, which is *not* a function of the observed **y**, but harder to get.
- The latter approach is harder to implement, but conveniently yields $\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}) \approx [-E\{D^2 L(\boldsymbol{\beta})\}]^{-1}$ evaluated at $\hat{\boldsymbol{\beta}}$ when the process is done.

Newton-Raphson for Logistic Regression

$$\begin{aligned}
\boldsymbol{\beta}_{new} &= \boldsymbol{\beta}_{old} - \left(\frac{\partial^2 \ell}{\partial^2 \boldsymbol{\beta}}\right)^{-1}\left(\frac{\partial \ell}{\partial \boldsymbol{\beta}}\right) \\
\boldsymbol{\beta}_{new} &= \boldsymbol{\beta}_{old} + (\mathbf{X}^T W X)^{-1} X^T (y - p) \\
\boldsymbol{\beta}_{new} &= (X^T W X)^{-1} X^T W [X \boldsymbol{\beta}_{old} + W^{-1}(y - p)] \\
\boldsymbol{\beta}_{new} &= (X^T W X)^{-1} X^T W z,
\end{aligned}$$

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

- if z is viewed as a response and $\mathbf{X}$ is the input matrix, $\boldsymbol{\beta}_{new}$ is the solution to a weighted least square problem.

$$\boldsymbol{\beta}_{new} = \text{argmin}_{\boldsymbol{\beta}} (z - \mathbf{X}\beta)^T W (z - \mathbf{X}\beta)$$

- z is referred to as the adjusted response.
- The algorithm is referred to as iteratively reweighted least square (IRLS)

Iteratively Re-weighted Least Squares (IRLS)

To set up the Newton-Raphson

- Set $\boldsymbol{\beta}$ to some initial value
- Set threshold values ϵ for convergence
- Set an iteration counter to track the number of iterations.

Iterratively Re-weighted Least Squares (IRLS)

- Set β to its initial value, $\beta_0 = \log\left(\frac{\bar{y}}{1-\bar{y}}\right)$
- Calculate p using $p = \frac{e^{\mathbf{X}\beta}}{1+e^{\mathbf{X}\beta}}$
- Calculate W using the updated p.
- Calculate $z = \mathbf{X}\beta + W^{-1}(y - p)$
- Update $\beta = (X^T W X)^{-1} X^T W z$
- Check if $|\beta_{new} - \beta_{old}| < \epsilon_1$, and $f(\beta_{old}) - f(\beta_{new}) < \epsilon_2$

Notice that in logistic regression $E\{D^2 L(\beta)\} = D^2 L(\beta)$, hence Newton-Raphson (NR) and Fisher Scoring methods ($E\{D^2 L(\beta)\}$) are equivalent. For other models, there is a difference between NR and Fisher Scoring. Many statistical packages such as SAS, R use Fisher Scoring as default.

Logistic Regression Inference

- The resulting estimate is consistent and it's large-sample variance is

$$\text{var}(\widehat{\beta}) = (X^T W X)^{-1}$$

- The Wald test for testing individual regression coefficient: $H_0 : \beta_i = 0$ versus $H_a : \beta_i \neq 0$ can be written as:

$$Z = \frac{\widehat{\beta}_i}{SE(\widehat{\beta}_i)}$$

- The $(1 - \alpha)\%$ confidence interval can be constructed as

$$\widehat{\beta}_i \pm Z_{1-\alpha/2} SE(\widehat{\beta}_i)$$

General Remarks

- There is an extensive literature on conditions for existence and uniqueness of MLEs for logistic regression
- MLEs may not exist. One case is when the data has "separation" of covariates (e.g., all success to left and all failures to right for some value of x.)