Unregularized and regularized logistic regression

In a Generalized Linear Model, we can express the distribution for \( y \) in the canonical form, which is
\[
f(y; \theta) = \exp\left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\},
\]
where \( \theta \) is called canonical parameter and \( \phi \) is called the dispersion parameter. And the log-likelihood is
\[
l(y; \theta) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi).
\]
We can obtain the first derivative or score of the log-likelihood w.r.t \( \theta \), which is
\[
l'(\theta; y) = \frac{y - b'(\theta)}{a(\phi)},
\]
and the second derivative, which is
\[
l''(\theta; y) = -\frac{b''(\theta)}{a(\phi)}.
\]

By the known property, \( E[l'(\theta; y)] = 0 \), it follows that \( \mu = b'(\theta) \). The information equality shows
\[
\text{var}[l'(\theta; y)] = E[l'^2(\theta; y)] - E^2[l'(\theta; y)] = -E[l''(\theta; y)]
\]
\[
\Rightarrow E[l'^2(\theta; y)] = -E[l''(\theta; y)],
\]
therefore it follows \( E\left[ \frac{(y-\mu)^2}{a^2(\phi)} \right] = \frac{b''(\theta)}{a(\phi)} \). Thus, we have \( \text{var}(y) = E[(y - \mu)^2] = b''(\theta)a(\phi) \).

For the canonical form of the GLM, we have the linear predictor
\[
\eta = \beta^T x,
\]
link function

\[ g(\mu) = \eta, \]

and the obtained fact

\[ \mu = b'(\theta). \]

Now we can establish the Fisher Scoring algorithm for the GLM model:

\[ \beta^{(t+1)} = \beta^{(t)} + (-E[l''(\beta^{(t)})])^{-1}l'(\beta^{(t)}), \]

where \( l'(\beta^{(t)}) \) is the score and \( -E[l''(\beta^{(t)})] \) is the expected information. \( \forall \beta_j \), we have the chain equation:

\[ \frac{\partial l}{\partial \beta_j} = \frac{\partial l}{\partial \theta} \left( \frac{\partial \theta}{\partial \mu} \right) \left( \frac{\partial \mu}{\partial \eta} \right) \left( \frac{\partial \eta}{\partial \beta_j} \right). \]

And we can derive the expression of them with ease:

\[ \frac{\partial l}{\partial \theta} = l'(\theta; y) = \frac{y - b'(\theta)}{a(\phi)}, \]

\[ \frac{\partial \theta}{\partial \mu} = 1, \quad \frac{\partial \mu}{\partial \theta} = b''(\theta) = \frac{a(\phi)}{\text{var}(y)}. \]

\[ \frac{\partial \eta}{\partial \beta_j} = x_{ij}, \forall x_i. \]

Combining the above results, we have

\[ \frac{\partial l}{\partial \beta_j} = \frac{(y - \mu)}{\text{var}(y)} \left( \frac{\partial \mu}{\partial \eta} \right) x_{ij}, \quad (1) \]

and by using the property of information equality and the derived results above, we have

\[ -E\left[ \frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right] = E\left[ \left( \frac{\partial l}{\partial \beta_j} \right) \left( \frac{\partial l}{\partial \beta_k} \right) \right] \]

\[ = E\left[ \left( \frac{y - \mu}{\text{var}(y)} \right)^2 \left( \frac{\partial \mu}{\partial \eta} \right)^2 x_{ij} x_{ik} \right] \]

\[ = \frac{1}{\text{var}(y)} \left( \frac{\partial \mu}{\partial \eta} \right)^2 x_{ij} x_{ik}. \]

(2)

With equations 1 and 2, we can construct the Fisher scoring algorithm. By rewriting equation 1 in the vector form, we obtain

\[ \frac{\partial l}{\partial \beta} = X^T A(y - \mu), \]

where

\[ X = (x_1, \ldots, x_N)^T, \]

\[ 2 \]
\[
A = \text{diag}\{\text{var}(y_i)^{-1}(\frac{\partial \mu_i}{\partial \eta_i})\},
\]
\[
y = (y_1, \ldots, y_N),
\]
and
\[
\mu = (\mu_1, \ldots, \mu_N).
\]
Similarly, we have
\[
-\mathbb{E}[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}] = X^TWX.
\]
where
\[
W = \text{diag}\{w_i\}
\]
\[
= \text{diag}\{\text{var}(y_i)^{-1}(\frac{\partial \mu_i}{\partial \eta_i})^2\}
\]
\[
= \text{diag}\{\text{var}(y_i)(\frac{\partial \eta_i}{\partial \mu_i})^2)^{-1}\}.
\]
Hence, we can construct the Fisher scoring as
\[
\beta^{(t+1)} = \beta^{(t)} + \{-\mathbb{E}[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}]\}^{-1} \frac{\partial l}{\partial \beta_j} \Rightarrow
\]
\[
\beta^{(t+1)} = \beta^{(t)} + (X^TWX)^{-1}X^TA(y - \mu).
\]
(3)
By rewriting equation 3, we have
\[
\beta^{(t+1)} = (X^TWX)^{-1}[X^TWX\beta^{(t)} + X^TWA(y - \mu)].
\]
(4)
Since \(\eta = \beta^T x\), we can write
\[
X\beta = (\eta_1, \ldots, \eta_N)^T = \eta.
\]
And
\[
A = W(\frac{\partial \eta}{\partial \mu}),
\]
where \(\frac{\partial \eta}{\partial \mu} = \text{diag}(\frac{\partial \eta_i}{\partial \mu_i}).\) Hence, we can write equation 4 as
\[
\beta^{(t+1)} = (X^TWX)^{-1}X^TWz,
\]
(5)
where
\[
z = \eta + (\frac{\partial \eta}{\partial \mu})(y - \mu) = (z_1, \ldots, z_N)^T,
\]
and elementwisely,
\[
z_i = \eta_i + (\frac{\partial \eta_i}{\partial \mu_i})(y_i - \mu_i),
\]
(6)
and
\[ w_i = \text{var}(y_i) \left( \frac{\partial \eta_i}{\partial \mu_i} \right)^2 \].

(7)

In the logistic regression, letting \( x_0 = 1 \), then we can write \( \beta_0 + x_i^T \beta \) as \( \beta^T x_i \). For each individual trial, we have the canonical form
\[
\begin{aligned}
f(y) &= \exp\{\log\{\frac{\exp(\beta^T x)}{1 + \exp(\beta^T x)}\}\}^{y} \left( \frac{1}{1 + \exp(\beta^T x)} \right)^{1-y} \\
&= \exp\{y\beta^T x - \log[1 + \exp(\beta^T x)]\},
\end{aligned}
\]
and the log-likelihood is
\[
l = y\beta^T x - \log[1 + \exp(\beta^T x)].
\]

Hence, we can write the corresponding components of the canonical form:
\[
\begin{aligned}
y &= y \\
\theta &= \beta^T x \\
a(\phi) &= 1 \\
b(\theta) &= \log\{1 + \exp(\beta^T x)\} \\
c(y, \phi) &= 0.
\end{aligned}
\]

By easy derivation, we can also obtain the following relations:
\[
\begin{aligned}
E[y] &= \mu = b'(\theta) = \frac{\exp(\theta)}{1 + \exp(\theta)} = \pi \\
\text{var}(y) &= a(\phi)b''(\theta) = b''(\theta) = \frac{\exp(\theta)}{1 + \exp(\theta)} \frac{1}{1 + \exp(\theta)} = \pi(1 - \pi) = \mu(1 - \mu) \\
\eta &= g(\mu) = \text{logit}(\pi) = \log\left( \frac{\pi}{1 - \pi} \right) = \beta^T x, \text{ where } \pi = Pr(G = 1|\mathbf{x}) = \frac{\exp(\theta)}{1 + \exp(\theta)}
\end{aligned}
\]
Since \( \mu = \pi \),
\[
\eta = \log\left( \frac{\mu}{1 - \mu} \right) = -\log(\mu^{-1} - 1).
\]
Therefore, we can obtain
\[
\frac{\partial \eta}{\partial \mu} = \mu^{-1} = \frac{1}{(1 - \mu)\mu} = \frac{1}{p(\mathbf{x})(1 - p(\mathbf{x}))},
\]
where \( p(\mathbf{x}) = \pi = Pr(G = 1|\mathbf{x}) \). Plugging back to equations 6 and 7, we get
\[
z_i = \beta^T x + \frac{y - p(\mathbf{x}_i)}{p(\mathbf{x}_i)(1 - p(\mathbf{x}_i))}
\]
and
\[
    w_i = \left[ \text{var}(y_i) \left( \frac{\partial \eta_i}{\partial \mu_i} \right)^2 \right]^{-1}
    = \left[ \pi(1 - \pi) \left( \frac{1}{\pi(1 - \pi)} \right)^2 \right]^{-1}
    = \pi(1 - \pi)
    = p(x_i)(1 - p(x_i)).
\]

Evaluating at the current parameter \( \tilde{\beta} \), we can obtain
\[
    z_i = \tilde{\beta}^T x_i + \frac{y_i - \tilde{p}(x_i)}{\tilde{p}(x_i)(1 - \tilde{p}(x_i))}
\]
as the working variate, and
\[
    w_i = \left[ \text{var}(y_i) \left( \frac{\partial \eta_i}{\partial \mu_i} \right)^2 \right]^{-1}
    = \tilde{p}(x_i)(1 - \tilde{p}(x_i))
\]
as the weight, which are corresponding to the content in the paper.

From the Equation 15 in the paper, which is the form of \( l_Q \), we can have the same algorithm by constructing the Fisher Scoring, by using the Newton–Raphson method first. In this case, the iteration can be formed in this way:
\[
    \beta^{(t+1)} = \beta^{(t)} + \frac{l_Q'}{l_Q},
\]
and assyptotically, this is equivalent to
\[
    \beta^{(t+1)} = \beta^{(t)} + \frac{l_Q'}{E[l_Q^2]},
\]
which is the Fisher Scoring. A easy rearrange of the terms will lead this equivalent to equation 5, by using the given \( z_i \) and \( w_i \) in the paper.

Putting all the above together, we have the IRWLS update algorithm to find the MLE in a logistic regression model:
\[
    \beta^{(t+1)} = (X^T WX)^{-1} X^T W z,
\]
where \( z = (z_1, \ldots, z_i, \ldots, z_N) \), and \( W = \text{diag}(w_i) \). Also, theoretically, a cholesky decomposition of \( X^T WX \) that is s.p.d to \( LL^T \), instead of computing the inverse, can accelerate the speed, by forming \( LL^T \beta^{(t+1)} = X^T W z \).