Exercise 1: Preliminaries **Cenying(Tracy) Yang**
SDS 385 Due Date: 8/29/2016
Prof. James Scott

Linear Regression

(A) **Solution**
WLS objective function in terms of vectors and matrices:
$\min_{\beta} f(\beta) \doteq \frac{1}{2}(y - X\beta)^T W(y - X\beta)$
By first order condition, we have
$\nabla f(\beta) = 0 \Rightarrow X^T W(y - X\beta) = 0 \Rightarrow (X^T W X)\hat{\beta} = X^T W y$,
where W is the diagonal matrix of weights.

Generalized linear models

(A) **Solution**
Since $y_i \sim Bin(m_i, w_i), p(y_i \mid \beta) = {}^nC_k w_i^{y_i}(1 - w_i)^{m_i - y_i}$.
Taking log, we have $\log p(y_i \mid \beta) = k + y_i \log w_i + (m_i - y_i)\log(1 - w_i)$, where k is a constant.
Plug into negative log likelihood, we have

$$\begin{aligned} l(\beta) &= -\log\{\prod_{i=1}^{N} p(y_i \mid \beta)\} \\ &= -\sum_{i=1}^{N} \log p(y_i \mid \beta) \\ &= -\sum_{i=1}^{N} y_i \log w_i + (m_i - y_i)\log(1 - w_i) \end{aligned} \tag{1}$$

where we omit the constant term k since it does not include parameters β, and hence does not affect optimization.
Denote $w_i \doteq h_\beta(x_i) \doteq g(x_i^T\beta) = \frac{1}{1+e^{-x_i^T\beta}}$, where $g(z) = \frac{1}{1+e^{-z}}$. Its derivative is

$$\begin{aligned} g^{'}(z) &= \frac{d}{dz}\frac{1}{1+e^{-z}} \\ &= \frac{1}{(1+e^{-z})^2}(e^{-z}) \\ &= \frac{1}{(1+e^{-z})}(1 - \frac{1}{(1+e^{-z})}) \\ &= g(z)(1 - g(z)) \end{aligned} \tag{2}$$

Denote $l_i(\beta) \doteq y_i \log w_i + (m_i - y_i)\log(1-w_i)$. By (1), we have $l(\beta) = -\sum_{i=1}^{N} l_i(\beta)$. Taking gradient of $l_i(\beta)$, we have

$$
\begin{aligned}
\nabla l_i(\beta) &= (y_i \frac{1}{g(x_i^T\beta)} - (m_i - y_i)\frac{1}{1-g(x_i^T\beta)})\nabla g(x_i^T\beta) \\
&= (y_i \frac{1}{g(x_i^T\beta)} - (m_i - y_i)\frac{1}{1-g(x_i^T\beta)}) g(x_i^T\beta)(1-g(x_i^T\beta))\nabla x_i^T\beta \\
&= (y_i(1-g(x_i^T\beta)) - (1-y_i)g(x_i^T\beta))x_i \\
&= (y_i - m_i w_i)x_i \\
&= (y_i - \frac{m_i}{1+e^{-x_i^T\beta}})x_i
\end{aligned} \tag{3}
$$

The second equality is by (2). The gradient of $l(\beta)$ is as follows

$$
\begin{aligned}
\nabla l(\beta) &= -\sum_{i=1}^{N} \nabla l_i(\beta) \\
&= -\sum_{i=1}^{N} (y_i - \frac{m_i}{1+e^{-x_i^T\beta}})x_i
\end{aligned} \tag{4}
$$

(C) **Solution**

The second order Taylor approximation of $l(\beta)$ at β_0 expands as follows

$$
l(\beta) = l(\beta_0) + (\beta-\beta_0)^T \nabla l(\beta_0) + \frac{1}{2}(\beta-\beta_0)^T H(l)(\beta)(\beta-\beta_0) \tag{5}
$$

As defined before, $g(x_i^T\beta) = \frac{1}{1+e^{-x_i^T\beta}}$. Then, the gradient $\nabla l(\beta_0)$ takes the form of

$$
\nabla l(\beta_0) = X^T(m_i g(X\beta) - y) \tag{6}
$$

The Hessian $H(l)(\beta)$ takes the form of

$$
\begin{aligned}
H(l)(\beta) &= X^T diag(m_i g(x_i^T\beta)(1-g(x_i^T\beta)))X \\
&= X^T D X
\end{aligned} \tag{7}
$$

where D is a diagonal matrix with $D_{ii} = m_i g_i(x_i^T\beta)(1-g_i(x_i^T\beta))$, where $g_i(x_i^T\beta) = \frac{1}{1+e^{-x_i^T\beta_i}}$. Plug equation (6) and (7) into (5), we get

$$
\begin{aligned}
l(\beta) &= l(\beta_0) + (\beta-\beta_0)^T X^T(m_i g(X^T\beta) - y) + \frac{1}{2}(\beta-\beta_0)^T X^T D X(\beta-\beta_0) \\
&= l(\beta_0) + (X\beta - X\beta_0)^T(m_i g(X^T\beta) - y) + \frac{1}{2}(X\beta - X\beta_0)^T D(X\beta - X\beta_0) \\
&= l(\beta_0) + (X\beta - X\beta_0)^T(m_i g(X^T\beta) - y) + \frac{1}{2}(X\beta_0 - X\beta)^T D(X\beta_0 - X\beta) \\
&= l(\beta_0) + (X\beta - X\beta_0)^T(m_i g(X^T\beta) - y) + \frac{1}{2}((X\beta_0 - X\beta)^T D((X\beta_0 - X\beta)
\end{aligned} \tag{8}
$$

We want to transform it into the form of $\frac{1}{2}((z - X\beta)^T W((z-X\beta) + c$. Here is the trick. Take a quadratic equation for a vector x, $a + b^T x + \frac{1}{2}x^T C x$, and we want to convert it into the form $\frac{1}{2}(x-m)^T M(x-m) + v$. Assume C is symmetric, then $M = C$, $m = -C^{-1}b$, $v = a - \frac{1}{2}b^T C^{-1} b$. Therefore, in our case, we have

- $W = D$
- $z - X\beta = -D^{-1}(m_i g(X^T\beta) - y)) \implies z = -D^{-1}(m_i g(X^T\beta) - y)) + X\beta$