IRWLS for LOGIT regression

In a logit (logistic) regression, the goal is the estimate of the beta coefficients. The problem can be solved with the IRLS algorithm by using the *Newton-Raphson method*.

As you already know, the Newton-Raphson method is a powerful technique for solving equations numerically. Like so much of the differential calculus, it is based on the simple idea of linear approximation.

The (conditional) likelihood function of the logit regression is, for response 0-1,

$$\prod_{i=1}^{n} p(x;\beta)^{y_i}(1-p(x;\beta))^{1-y_i}, \tag{1}$$

as well as, for grouped data it is

$$\prod_{i=1}^{n} p(x;\beta)^{y_i}(1-p(x;\beta))^{m_i-y_i}. \tag{2}$$

Note that they are exactly the same, placing $m_i = 1$ in the first case.

The conditional log-likelihood is instead (response 0-1):

$$\begin{aligned} L(\beta) &= \sum_{i=1}^{n} \left[y_i ln P(x_i;\beta) + (1-y_i) ln\left(1-P(x_i;\beta)\right)\right] \\ &= \sum_{i=1}^{n} \left[ln(1-P(x_i;\beta)) + y_i ln\left(\frac{P(x_i;\beta)}{1-P(x_i;\beta)}\right)\right], \end{aligned}$$

while for grouped data is

$$\begin{aligned} L(\beta) &= \sum_{i=1}^{n} \left[y_i ln P(x_i;\beta) + (m_i-y_i) ln(m_i - P(x_i;\beta))\right] = \\ &= \sum_{i=1}^{n} m_i ln(1-P(x_i;\beta)) + y_i ln\left(\frac{P(x_i;\beta)}{1-P(x_i;\beta)}\right). \end{aligned}$$

Of course, $P(x_i;\beta) = \frac{exp(x_i\beta)}{1+exp(x_i\beta)}$ and $1-P(x_i;\beta) = \frac{1}{1+exp(x_i\beta)}$, hence we can

write the log-likelihood in terms of β as

$$\begin{aligned} L(\beta) &= \sum_{i=1}^{n} m_i ln\left(\frac{1}{1+exp(x_i\beta)}\right) + y_i ln\left(\frac{\frac{exp(x_i\beta)}{1+exp(x_i\beta)}}{\frac{1}{1+exp(x_i\beta)}}\right) \\ &= \sum_{i=1}^{n} m_i ln\left(\frac{1}{1+exp(x_i\beta)}\right) + y_i(x_i\beta) \\ &= \sum_{i=1}^{n} y_i(x_i\beta) - m_i ln(1+exp(x_i\beta))\,. \end{aligned}$$

For 0-1 loss, set $m_i = 1$, of course.

Moreover, we have:

$$\frac{\partial L(\beta)}{\beta_j} = \sum_{i_1}^{n} x_{ij}(y_i - m_i p_i(x_i;\beta)) \quad j = 1,\ldots,p \text{ with p=num. of predictors,}$$

and

$$\frac{\partial^2 L(\beta)}{\partial\beta_j\partial\beta_k} = -\sum_{i_1}^{n} x_{ij}x_{ik}m_i\left(p_i(x_i;\beta)(1-p_i(x_i;\beta))\right).$$

The IRLS algorithm works in this way:

- Set β_0 as initial vector of parameters
- Set $P = exp(X\beta_0)/(1+exp(X\beta_0))$

1. Set W as a *diagonal weight matrix* of size $n \times n$ whose elements are $m_i P_i(1-P_i)$
2. $\beta^+ = \beta_0 + (X'WX)^{-1}X'(Y - m \times P) = (X'WX)^{-1}X'Wz$, with $z = X\beta_0 + W^{-1}(y - m \times P)$
3. Set $\beta^+ = \beta_0$

Operatively:

1. set $\beta_0 = 0$
2. compute $P(X;\beta_0) = exp(X\beta_0)/(1+exp(X\beta_0))$
3. compute W
4. $z = X\beta_0 + W^{-1}(y - mP)$
5. $\beta^+ = (X'WX)^{-1}X'Wz$
6. stop if $\beta^+ = \beta_0$

Don't forget to insert a column of ones in the design matrix. The variance of the regression coefficients is equal to the diagonal elements of the matrix $(X'WX)^{-1}$. Implement in either R or Python (or both) the IRWLS algorithm that returns regression coefficients, standard errors and t-ratio of the coefficients.

Consider Aids2 data (package MASS). Write a program for estimating the logistic regression parameters. Compare the estimated beta coefficients coming from your code with the ones estimated by the function glm(, (binomial)). Regress "status" on "age" and "sex". If necessary, convert labels (for status and sex) into numbers.

Next, find by yourself a data set with grouped data (or, if you want, aggregate data from Aids2 data). Then, re-estimate the coefficients and standard errors. Compute the log-likelihood at the convergence point ($loglik = -2(\sum_i y_i ln(P_i) + (m_i - y_i) ln(1 - P_i))$). There is some difference?

Remember that the IRWLS remains the same for both grouped and individual data: just set $m_i = 1$ in the second case

Please send your homework by next December 23.