GLM estimation and model fitting

Patrick Breheny

February 19
Introduction

- In previous lectures, we’ve discussed the theoretical properties of $\hat{\beta}$, the regression coefficients of a generalized linear model.
- We turn our attention today to a more practical matter: how do we actually solve for $\hat{\beta}$?
- This is a more challenging question than it sounds – in general, there is no closed form solutions for the maximum likelihood estimator $\hat{\beta}$.
- Nevertheless, it turns out that we can combine the ideas from our last two lectures (Taylor series approximations and iteratively reweighted least squares) to obtain an algorithm for obtaining $\hat{\beta}$. 
MLEs for GLMs

- As we have discussed previously, we obtain MLEs by setting the score vector equal to 0.
- Recall that for a GLM using the canonical link function, the score vector is
  \[ u(\beta) = \phi^{-1}X^T(y - \mu) \]
- Note that in the above equation, \( \mu \) is a function of \( \eta = X\beta \); however, it need not be a linear function, and if it is not, we lack a closed-form solution for \( \beta \).
Taylor approximation for $\mu$

Nevertheless, we can apply a Taylor series approach to obtain the following approximation about the point $\tilde{\beta}$:

$$\mu \approx \tilde{\mu} + W (X\beta - X\tilde{\beta}),$$

where $\tilde{\mu} = g^{-1}(X\tilde{\beta})$

Note that the above result rests on the following proposition

**Proposition:** If $g$ is the canonical link, then

$$\frac{d}{d\eta} g^{-1}(\eta) = W(\eta)$$
Main result

- Thus, we obtain the following linear approximation to the score for $\beta$:

$$
\frac{\partial \ell}{\partial \beta} \approx \phi^{-1} X^T W (z - X \beta),
$$

where $z = X \tilde{\beta} + W^{-1} (y - \tilde{\mu})$ is known as the adjusted response.

- Note that this approximation is based at $\tilde{\beta}$ or, equivalently, $\tilde{\mu}$, which are treated as constants in the above expression, thereby rendering the score equation linear in $\beta$ after the approximation.

- Again, recall that this approximation will be accurate near the fitted values $\tilde{\mu}$, but not necessarily accurate far away from them.
As we saw previously, this gives the maximum likelihood estimate

\[ \hat{\beta}^{(m)} = (X^T WX)^{-1} X^T W z \]

Note that \( W \) here plays the role of the weights in weighted least squares, and for that reason is often referred to as the 
\textit{weight matrix}.

Again, recall that for the canonical link, \( W \) is entirely determined by the mean-variance relationship, and that it plays a prominent role in the variability of \( \hat{\beta} \) as well.

Note that in the above equation, we require a superscript on \( \hat{\beta}^{(m)} \) because this is a case of unknown weights, where \( W \) (and \( z \)) will change depending on \( \hat{\beta} \) and vice versa.
As we saw earlier, one way to address this problem is to iterate the process of reweight–estimate–reweight–estimate–... until convergence; this \textit{iteratively reweighted least squares} (IRLS) algorithm is how generalized linear models are fit:

1. Choose an initial value $\hat{\beta}^{(0)}$
2. For $m = 0, 1, 2, \ldots$,
   a. Calculate $z$ and $W$ based on $\hat{\beta}^{(m)}$
   b. Solve for $\hat{\beta}^{(m+1)}$
   c. Check to see whether $\hat{\beta}$ has converged; if yes, then stop
The IRLS algorithm for GLMs

Unique solutions?

The Newton-Raphson algorithm

This IRLS algorithm is a special case of a more general approach to optimization called the *Newton-Raphson* algorithm.

The Newton-Raphson algorithm calculates iterative updates via

\[ \hat{\beta}^{(m+1)} = \hat{\beta}^{(m)} - H^{-1}u, \]

where \( u \) is the score vector and \( H \) is the Hessian matrix (the first and second derivatives of the log-likelihood, respectively), both of which are evaluated at \( \hat{\beta}^{(m)} \).

It can be shown (homework) that this produces the same iterative updates as IRLS.
Unique solutions and rank

- Recall that, for linear regression, a full-rank design matrix $X$ implied that there was exactly one unique solution $\hat{\beta}$ which minimized the residual sum of squares.
- A similar result holds for generalized linear models: if $X$ is not full rank, then there is no unique solution which maximizes the likelihood.
However, two additional issues arise in generalized linear models:

- Although a unique solution exists, the IRLS algorithm is not guaranteed to find it.
- It is possible for the unique solution to be infinite, in which case the estimates are not particularly useful and inference breaks down.

The first issue is uncommon; we will see an example of the second issue in an upcoming lecture.