Generalized linear models: Estimation and model fitting

Patrick Breheny

March 22

Introduction

- We've discussed the ingredients that go into specifying a generalized linear model
- In this lecture, we address the question: how do we actually estimate the regression coefficients?

Taylor series approximations

- In generalized linear models, both model fitting (today) and inference (next lecture) rely heavily on making linear/quadratic *Taylor series approximations*
- Suppose that $f(x)$ is a differentiable function, but is not necessarily linear and possibly rather complicated
- A simple approximation to $f(x)$, valid in the neighborhood of a point x_0, is given by

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

note that this function is linear in x

Second-order approximations

- A more complicated, but more accurate, approximation is given by

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2;$$

  a quadratic function in x
- We could keep going, of course, to higher and higher orders, but in practice, first and second orders usually suffice
- Taylor's theorem guarantees that any sufficiently smooth function can be approximated in this way, and provides bounds for the error of the approximation

Multidimensional approximations

- Taylor series approximations can be conducted in higher dimensions as well:

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla^T(\mathbf{x} - \mathbf{x}_0),$$
$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla^T(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T\mathbf{H}(\mathbf{x} - \mathbf{x}_0),$$

  where $\nabla = \frac{\partial f}{\partial \mathbf{x}}$ and $\mathbf{H} = \frac{\partial^2 f}{\partial \mathbf{x}^2}$

- ∇ is sometimes referred to as the *gradient* and $\mathbf{H}$ as the *Hessian*

Introduction

- With those preliminaries out of the way, we are now in a position to estimate the regression coefficients
- As we mentioned earlier, the reason for restricting ourselves to the exponential family is that it facilitates maximum likelihood estimation
- Unfortunately, we cannot, in general, obtain a closed form solution for the maximum likelihood estimator
- However, after making a Taylor series approximation to the likelihood about the fitted values $\hat{\boldsymbol{\mu}}$, we obtain an estimator that is equivalent to the weighted least squares estimate

Main result

- Specifically, suppose we are taking a Taylor series approximation about the fitted values $\tilde{\boldsymbol{\mu}}$ resulting from regression coefficients $\tilde{\boldsymbol{\beta}}$; then

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

  where $\mathbf{W}$ is a diagonal matrix with elements $\{1/g'(\mu_i)\}$ and $\mathbf{z} = \mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{W}^{-1}(\mathbf{y} - \tilde{\boldsymbol{\mu}})$ ($\mathbf{z}$ is sometimes referred to as the *adjusted response*)
- As a clarification, the value $\tilde{\boldsymbol{\beta}}$ used to make the approximation is treated as a constant in the above expression; $\boldsymbol{\beta}$ is the only variable, and the score equation is linear in $\boldsymbol{\beta}$ after the approximation

Iteration

- As we saw previously, this gives the maximum likelihood estimate

$$\widehat{\boldsymbol{\beta}}^{(m)} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z}$$

- The superscript on $\widehat{\boldsymbol{\beta}}^{(m)}$ is because this is a case of unknown weights, where $\mathbf{W}$ (and $\mathbf{z}$) will change depending on $\widehat{\boldsymbol{\beta}}$ 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 *iteratively reweighted least squares* (IRLS) algorithm is how generalized linear models are fit

IRLS algorithm: summary

In summary, then, the algorithm goes like this:

(1) Choose an initial value $\widehat{\boldsymbol{\beta}}^{(0)}$

(2) For $m = 0, 1, 2, \ldots$,

  (a) Calculate $\mathbf{z}$ and $\mathbf{W}$ based on $\widehat{\boldsymbol{\beta}}^{(m)}$

  (b) Solve for $\widehat{\boldsymbol{\beta}}^{(m+1)}$

(3) Repeat step (2) until convergence

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

$$\widehat{\boldsymbol{\beta}}^{(m+1)} = \widehat{\boldsymbol{\beta}}^{(m)} - \mathbf{H}^{-1}\mathbf{u},$$

  where $\mathbf{u}$ is the score vector and $\mathbf{H}$ is the Hessian matrix (the first and second derivatives of the log-likelihood, respectively), both of which are evaluated at $\widehat{\boldsymbol{\beta}}^{(m)}$
- It can be shown (homework) that this produces the same iterative updates as IRLS

Weights for the canonical link

- **Proposition:** If g is the canonical link for the exponential family, then $1/g'(\mu_i) = V(\mu_i)$.
- In other words, $\mathbf{W} = \mathbf{V}$, where $\mathbf{V}$ is a diagonal matrix with elements $\{V(\mu_i)\}$
- The weight matrix $\mathbf{W}$ plays a prominent role in inference as well; this proposition tells us that for the canonical link, $\mathbf{W}$ is entirely determined by the mean-variance relationship

Unique solutions and rank

- Recall that, for linear regression, $\mathbf{X}$ full rank implied that there was exactly one unique solution $\widehat{\beta}$ which minimized the residual sum of squares
- A similar result holds for generalized linear models: if $\mathbf{X}$ is not full rank, then there is no unique solution which maximizes the likelihood

Additional issues for GLMs

- 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 an example of the second issue in an upcoming lab