# A Little Optimization

Rob McCulloch

December 1, 2019

Logistic Regresssion

Newton's Method

Logit Log-Likelihood Derivatives: The Logit MLE

Iteratively Reweighted Least Squares

Optimization and Convexity/Concavity

Bayesian Posterior Mode

# Logistic Regresssion

Consider the linear regression model with iid normal errors:

$$Y = X\,\beta + \epsilon, \;\; \epsilon \sim N(0, \sigma^2 I)$$

We know how to get the least squares estimate of $\beta$ which is also the mle.

We solve the equations:

$$X'(Y - X\beta) = 0$$

We can solve these equation direclty using $\hat{\beta} = (X'X)^{-1}X'Y$ or use the cholesky decomposition of $X'X$ or the QR decomposition of $X$.

We have also learned how to do a Bayesian analysis of the linear model using the normal prior for $\beta$ and the inverted chi-squared prior for $\sigma$ and the Gibbs sampler:

$$\beta \,|\, \sigma, (Y, x), \;\; \sigma \,|\, \beta, (Y, X)$$

Perhaps after linear regression, the most fundamental model in applied statistics is logistic regression.

We want to use linear methods, but now are response is binary: $y \in \{0, 1\}$.

For a single $(x, y)$ observation we have

$$P(Y = 1 \,|\, x, \beta) = F(x' \beta), \;\; F(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)}.$$

*How do we compute the mle???*

*How do we do a Bayesian analysis*

# Newton's Method

Newton's method is a very basic method in optimization.

We will use it to compute the logit mle, and the Bayesian posterior mode.

Suppose $f : \beta \rightarrow R$, $\beta \in R^p$.

We want to minimize (or maximize) $f$.

We will need the first derivative:

$$f'(\beta) = \nabla f(\beta) = [\frac{\partial f(\beta)}{\partial \beta_1}, \frac{\partial f(\beta)}{\partial \beta_2}, \ldots, \frac{\partial f(\beta)}{\partial \beta_p}]$$

and the second derivative:

$$f''(\beta) = [\frac{\partial^2 f(\beta)}{\partial \beta_i \partial \beta_j}]$$

- the first derivative is called the gradient vector. My convention is that it is a $1 \times p$ row vector.
- the second derivative is a $p \times p$ symmetric matrix. It is called the Hessian.

## Newton's Method

Newton's method is iterative.

Let $\beta_i$ the value at iteration $i$.

- approximate f at $\beta_i$ by a quadratic using Taylors's theorem.
- optimize the quadratic: the solution is $\beta_{i+1}$.
- repeat until converged.

Taylor approximation:

$$f(\beta) \approx \tilde{f}(\beta) = f(\beta_i) + f'(\beta_i)(\beta - \beta_i) + \frac{1}{2}(\beta - \beta_i)' f''(\beta_i)(\beta - \beta_i)$$

Now to optimize the quadratic, we compute its gradient and set it equal to 0.

$$\nabla \tilde{f}(\beta) = f'(\beta_i) + (\beta - \beta_i)' f''(\beta_i)$$

We can solve $\nabla \tilde{f}(\beta) = 0$ with

$0 = f'(\beta_i) + (\beta - \beta_i)' f''(\beta_i)$

$-f'(\beta_i)[f''(\beta_i)]^{-1} = \beta' - \beta_i'$

$\beta' = \beta_i' - f'(\beta_i)[f''(\beta_i)]^{-1}$

$$\beta_{i+1} = \beta_i - [f''(\beta_i)]^{-1}[f'(\beta_i)]'$$

# Logit Log-Likelihood Derivatives: The Logit MLE

We will compute the first and second derivatives of the logit log likelihood.

First, we differentiate $F(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)}$:

$$F'(\eta) = \frac{(1 + e^{\eta})e^{\eta} - e^{\eta}e^{\eta}}{(1 + e^{\eta})^2} = F(\eta)(1 - F(\eta))$$

## The Likelihood

$$L(\beta) = \prod_{i=1}^{n} F(x_i'\beta)^{y_i}\,(1 - F(x_i'\beta))^{(1-y_i)}$$

Let $F_i = F(x_i'\beta)$.

$$\log L(\beta) = \sum y_i\,\log(F_i) + (1 - y_i)\log(1 - F_i)$$

$$\begin{aligned}
\log L'(\beta) &= \sum y_i x_i' \frac{F_i(1-F_i)}{F_i} + x_i'(1-y_i)[-\frac{F_i(1-F_i)}{(1-F_i)}] \\
&= \sum [y_i x_i'(1-F_i) - x_i'(1-y_i)F_i] \\
&= \sum x_i'(y_i - F_i) \\
&= (y - F)'X
\end{aligned}$$

$$\log L''(\beta) = -\sum x_i x_i' F_i(1 - F_i)$$
$$= -X'DX$$

where,

$$D = diag(F_i(1 - F_i))$$

So, to compute the logit mle:

$$
\begin{aligned}
\beta_{i+1} &= \beta_i - [-X'DX]^{-1}X'(y - F) \\
&= \beta_i + [X'DX]^{-1}X'(y - F)
\end{aligned}
$$

# Iteratively Reweighted Least Squares

Recall weighted least squares

$$Y = X\beta + \epsilon, \;\; \epsilon \sim N(0, \Sigma)$$

then,

$$\hat{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y$$

It may be helpful to rewrite the Newton iteration as a series of weighted regressions:

Let $\Sigma^{-1} = D$ and

$$Z = X\beta_i + D^{-1}(y - F)$$

then,

$$(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Z = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}(X\beta_i + D^{-1}(y - F))$$

$$= \beta_i + [X'DX]^{-1}X'(y - F)$$

Hence doing an iteratively (re)weighted least squares problem (IRLS) gets you the mle.

# Optimization and Convexity/Concavity

Recall that a function is convex if

$$f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha) f(x_2), \quad \alpha \in [0,1].$$

and concave if it goes the the other way,

$$f(\alpha x_1 + (1-\alpha)x_2) \geq \alpha f(x_1) + (1-\alpha) f(x_2), \quad \alpha \in [0,1].$$

Key:

If a function is convex then it has a unique global minimum and any local miniumum is the local minimum.

Same for concave and maximum.

Key:

If the Hessian is positive definite everywhere, then the function is convex.

Same for negative definite and concave.

$$a'[-X'DX]a = -v'Dv = -\sum v_i^2 F_i(1 - F_i) \leq 0.$$

Hence the logit logLikelihood is concave, hence Newton will converge to a global max.

# Bayesian Posterior Mode

We can use a very similar approach to compute the Bayesian posterior mode given a mulitvariate normal prior for $\beta$.

Let

$$p(\beta) \sim N(\bar{\beta}, A^{-1}).$$

Then,

$$p(\beta \,|\, X, Y) \propto L(\beta)\, p(\beta)$$

and

$$\log p(\beta \,|\, X, Y) = \log L(\beta) + log(p(\beta))$$

$$log(p(\beta)) = C - \frac{1}{2}(\beta - \bar{\beta})'A(\beta - \bar{\beta}) \equiv C + g(\beta)$$

$$g'(\beta) = -(\beta - \bar{\beta})A, \;\; g''(\beta) = -A.$$

So the Newton iterations become:

$$\beta_{i+1} = \beta_i + [X'DX + A]^{-1}[X'(y - F) - A(\beta_i - \bar{\beta})]$$

where $D$ and $F$ also depend on $\beta_i$.

Note:

Maximizing the log posteior is equivalent to minimizing

$$-\log L(\beta) + \frac{1}{2}(\beta - \bar{\beta})' A(\beta - \bar{\beta})$$

If we let $A = \lambda I$, $\bar{\beta} = 0$,and recall that $-\log L(\beta)$ is the deviance which is also called the cross-entropy loss then we minimize

$$Loss(y, \beta) + \lambda ||\beta||^2$$

Thus the Bayesian posterior mode can be viewed as an L2 regularized estimate of the coefficient vector.