LAPLACE'S METHOD AND A VARIATIONAL APPROXIMATION FOR A SMALL LOGISTIC REGRESSION EXAMPLE

PETER CARBONETTO*

Summary. The goal of these lecture notes is to develop a practical approach to implement hypothesis testing with logistic regression. We're going to look at two different approaches to this problem: using Laplace's method, and using variational approximation techniques.

Logistic regression. To make things concrete as possible, suppose we want to assess the ability of our variable (X) to predict, or explain, disease (Y). I denote sample i in our data set by x_i and y_i. The logistic model says that the logarithm of the odds of disease, denoted by log-odds(i), is a linear combination of x_i:

$$\text{log-odds}(i) = \log\left\{\frac{p(y_i = 1 \,|\, x_i, \beta)}{p(y_i = 0 \,|\, x_i, \beta)}\right\} = \beta_0 + x_i\beta.$$

I use π_i as shorthand for $p(y_i = 1 \,|\, x_i, \beta)$, and write the summation above using the dot product $x_i^T\beta$, so the logistic model is

$$\text{log-odds}(i) = \log\left\{\frac{\pi_i}{1 - \pi_i}\right\} = \beta_0 + x_i\beta.$$

In other words, the disease status is a coin toss with success rate

$$\pi_i = \sigma(x_i\beta),$$

where $\sigma(x) = 1/(1+e^{-x})$ is the sigmoid function (it is the inverse of the logit function).

Assuming independence of the samples, the likelihood of $y = (y_1, \ldots, y_n)^T$ given $x = (x_1, \ldots, y_n)^T$ and β is

$$p(y \,|\, x, \beta) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}.$$

One reason that the logistic model is so popular is because it is easy to compute the maximum likelihood or *maximum a posteriori* estimator. Instead, I'm going to focus on the problem of computing the ratio of the marginal likelihoods,

$$\frac{p(y \,|\, x, H_1)}{p(y \,|\, x, H_0)} = \frac{\int p(y \,|\, x, \beta)\, p(\beta)\, d\beta}{p(y \,|\, x, H_0)},$$

where H_0 is, for the sake of illustration, the null hypothesis that no factor contributes to disease risk, and H_1 is alternative hypothesis in which X increases or decreases susceptibility to disease. To compute this ratio of likelihoods—what we call the Bayes factor—we need to marginalize or integrate out the random vector β. This integral has no closed-form solution, so we must compute it *numerically*.

There is, of course, the option of using Monte Carlo methods, but I'm going to focus on a couple alternative approaches.

*Dept. of Human Genetics, University of Chicago, 920 E58th St., 4th floor, Chicago, IL, 60637

Laplace's method. A simple and general-purpose method we can use to approximate the marginal likelihood is Laplace's method, which is nothing more than a Taylor series approximation to the logarithm of the density function. Let's look at Laplace's method first in the general case.

For some nonlinear function $f(x)$, with $x \in \mathbb{R}^d$, suppose we want to compute the integral

$$I = \int e^{f(x)}\, dx$$

that has no known closed-form solution. (We'll write our marginal likelihood in this form.) First, we form a second-order Taylor series approximation about point $\hat{x}$:

$$f(x) \approx f(\hat{x}) + g^T(x - \hat{x}) + \tfrac{1}{2}(x - \hat{x})^T H(x - \hat{x}),$$

where $g = \nabla f(\hat{x})$ is the vector-valued function of first-order partial derivatives (the gradient) at $\hat{x}$, and $H = \nabla^2 f(\hat{x})$ is the matrix of second-order partial derivatives (the Hessian) at $\hat{x}$. This will be a good a approximation near $\hat{x}$, but will get increasingly worse as we get further away from $\hat{x}$.

The second-order Taylor-series approximation to the integral is

$$I \approx \int e^{f(\hat{x}) + g^T(x-\hat{x}) + \frac{1}{2}(x-\hat{x})^T H(x-\hat{x})}\, dx.$$

To solve this integral, we'll use the fact that the multivariate normal density with mean μ and covariance Σ integrates to one:

$$\int |2\pi\Sigma|^{-1/2} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}\, dx = 1.$$

Here, $|A|$ denotes the determinant of matrix A. Next, making the substitutions $\Sigma = -H^{-1}$ and $\mu = \Sigma g + \hat{x}$, and rearranging terms, we find that

$$I \approx |2\pi\Sigma|^{1/2} e^{f(\hat{x}) + \frac{1}{2}(\hat{x}-\mu)^T\Sigma^{-1}(\hat{x}-\mu)}.$$

Three remarks here. The first remark is that the covariance is equal to the negative of the inverse of the Hessian. Since, the covariance must be positive definite, the Hessian must be negative definite at $\hat{x}$. If $f(x)$ is a concave function, then the Hessian is guaranteed to be negative definite. This is the case for our logistic model.

Second, observe that this is general procedure; nowhere did we require that the function $f(x)$ correspond to a probability density function.

Third, if $f(x)$ corresponds to some density function, then what we have effectively done is replaced this density that is difficult to integrate with a normal density with mean μ and covariance Σ.

Before applying this to our problem, let me point out two special cases of Laplace's method. One way to improve the approximation is to first find a maximum of the function, so that the approximation is best where there is the most support. If $\hat{x}$ is a local maximum, it follows that $g = 0$, so the mean is simply $\mu = \hat{x}$, and the Laplace approximation reduces to

$$I \approx |2\pi\Sigma|^{1/2} e^{f(\hat{x})}.$$

The second interesting case occurs when we center our quadratic approximation at $\hat{x} = 0$. This might be a sensible course of action if we know that most of the risk factors will be zero (such as in a genome-wide association study). In this case,

$$I \approx |2\pi\Sigma|^{1/2} e^{f(\hat{x}) + \frac{1}{2}\mu^T\Sigma^{-1}\mu}.$$

Now let's apply this result to our marginal likelihood of interest. First, I need to rewrite our integral in the form

$$\int p(y \mid X, \beta)\, p(\beta)\, d\beta = C \int e^{f(\beta)}\, d\beta.$$

The function $f(\beta)$ plays the role of the log-density of the likelihood and the prior; $e^{f(\beta)}$ is the posterior up to a normalizing constant. Assuming the prior is normal with mean zero and variance σ_0^2, the likelihood times the prior is

$$p(y \mid X, \beta)\, p(\beta) = (2\pi\sigma_0^2)^{-1/2} e^{f(\beta)},$$

where

$$f(\beta) = \sum_{i=1}^{n} y_i \log \pi_i + \sum_{i=1}^{n} (1 - y_i) \log(1 - \pi_i) - \beta^2/(2\sigma_0^2).$$

Recall that β appears in the π_i's.

Notice that we can rewrite the log-density $f(\beta)$ as

$$f(\beta) = \sum_{i=1}^{n} (y_i - 1)\text{log-odds}(i) + \sum_{i=1}^{n} \log \pi_i - \beta^2/(2\sigma_0^2).$$

Here I've divided the terms in the exponent into two parts: those that are linear or quadratic in β, and those that are nonlinear in β. The only part that requires approximation is the nonlinear part.

Let's now investigate the first strategies I mentioned earlier: set $\hat{\beta}$ to the maximum of $f(\beta)$. To derive the result I'll use the fact that the gradient and Hessian of the log-density (*i.e.* the first and second-order derivatives with respect to β) work out to be

$$\begin{aligned} \nabla f(\beta) &= x^T (y - \pi) - \beta/\sigma_0^2 \\ \nabla^2 f(\beta) &= -1/\sigma_0^2 - x^T W x, \end{aligned}$$

where π is the vector with entries π_i, and W is the matrix with diagonal entries $\pi_i(1 - \pi_i)$. The expression for the gradient makes sense: ignoring the prior, we obtain a maximum, roughly speaking, when the predictions π_i match the observed labels y_i. For $\hat{\beta} = \beta^{(\text{MAP})}$, the final result is easy to calculate:

$$I \approx \sigma/\sigma_0 e^{f(\hat{\beta})},$$

with variance $\sigma^2 = (1/\sigma_0^2 + x^T W x)^{-1}$.

Variational method. The first step in developing our variational inference procedure is to deal with the nonlinear $\log \pi_i$ terms in our log-density function $f(\beta)$. The basic idea is to formulate a lower bound to the logarithm of the sigmoid function. Skipping the technical details (see Christopher Bishop's book), we obtain the lower bound

$$\log \sigma(x) \geq \log \sigma(\theta) + \tfrac{1}{2}(x - \theta) - \tfrac{u}{2}(x^2 - \theta^2),$$

where $u = \frac{1}{\theta}(\sigma(\theta) - \frac{1}{2})$, and $\theta \geq 0$ adjusts this lower bound. I will have one parameter θ for every sigmoid term or sample. Notice that the terms involving x—which will

later be replaced by linear combinations of β—are conveniently linear or quadratic. This leads directly to a lower bound on the log-likelihood $\log p(y \mid x, \beta)$, which I denote by $L(\beta; \theta)$:

$$L(\beta;\theta) = (y-1)^T \text{log-odds}(i) + \sum_{i=1}^{n} \log \sigma(\theta_i) + \tfrac{1}{2}(\text{log-odds}(i) - \theta_i) - \tfrac{u_i}{2}(\text{log-odds}(i)^2 - \theta_i^2).$$

When there is no intercept ($\beta_0 = 0$), this simplifies to

$$L(\beta;\theta) = \sum_{i=1}^{n} \log \sigma(\theta_i) + \tfrac{\theta_i}{2}(u_i\theta_i - 1) + (y - \tfrac{1}{2})^T x\beta - x^T U x\beta^2/2.$$

By extension, we obtain a lower bound on the marginal likelihood:

$$I = \int p(y \mid x, \beta)\, p(\beta)\, d\beta = \int p(\beta)\, e^{f(\beta)}\, d\beta \geq \int p(\beta)\, e^{L(\beta;\theta)}\, d\beta.$$

Here, U is the $n \times n$ matrix with diagonal entries u_i. Compare the lower bound $L(\beta; \theta)$ to the likelihood for a linear regression with coefficients β: $(y - \frac{1}{2})$ is a vector of measurements for a surrogate quantitative trait, and U scales the inverse of the covariance matrix.

Note: for the remainder, I'll assume $\beta_0 = 0$, just to make these expressions a little bit simpler. The MATLAB code allows for arbitrary intercept β_0.

Now that we've derived a lower bound on I, there are two pressing questions: How do we compute this lower bound? And how do we adjust the parameters $\theta = (\theta_1, \ldots, \theta_n)$ so that the lower bound is as tight as possible? To resolve the first question, suppose that the prior is normal with zero mean and variance σ_0^2. Then $p(\beta)\, e^{L(\beta;\theta)}$ is, up to a constant of proportionality, multivariate normal with mean $\mu = \sigma^2 x^T (y - \frac{1}{2})$ and variance $\sigma^2 = (1/\sigma_0^2 + x^T U x)^{-1}$. In this special case, the expression for the lower bound works out to be

$$I \geq \sigma/\sigma_0 \exp\left\{ \tfrac{1}{2}\mu^2/\sigma^2 + \textstyle\sum_{i=1}^{n} \log \sigma(\theta_i) + \tfrac{\theta_i}{2}(u_i\theta_i - 1) \right\}.$$

The second problem is perhaps most easily resolved by interpreting within the EM framework:in the E-step, compute expectations (the mean and covariance) of the unknowns β; and in the M-step, compute the *maximum a posteriori* estimator of θ, which amounts to maximizing the expected value of $L(\beta; \theta)$. (We can ignore the $\log p(\beta)$ term because θ does not affect the prior.)

To derive the M-step, first take the partial derivatives of the expected log-likelihood with respect to the variational parameters:

$$\frac{\partial E[L(\beta;\theta)]}{\partial \theta_i} = \frac{u_i'}{2}(\theta_i^2 - (x_i\mu)^2 - x_i^2\sigma^2),$$

where μ is the posterior mean and Σ is the posterior covariance computed in the E-step. The usual procedure is to set the partial derivatives to zero and solve for θ. At first glance, this does not appear to be possible. But a couple of observations will yield a closed-form solution: first, the slope u is symmetric in θ, so we only need to worry about the positive quadrant; second, for $\theta > 0$, u is strictly monotonic as a function of θ, and so u' is never zero. Therefore, we can solve for the fixed point:

$$\theta_i = \sqrt{(x_i^T\mu)^2 + x_i^T \Sigma x_i}.$$

This EM algorithm for adjusting the variational parameters θ has an intuitive appeal: we are adjusting the lower bound so that the approximation is tighest where there is most support for β.