Likelihood theory

Patrick Breheny

March 24

Introduction

- In today's lecture, we present the basic results of asymptotic likelihood theory, and show how these allow us to construct confidence intervals and carry out hypothesis tests for maximum likelihood estimates (such as regression coefficients in a GLM)
- We will be carrying out heuristic derivations of these results, rather than constructing rigorous asymptotic proofs
- To that end, we will use the symbol $\dot{\sim}$ to mean, "is approximately distributed as", or more specifically, that the use of this approximate distribution is justified asymptotically

Sampling distribution of the score statistic

- The standard result we will begin with is the sampling distribution of the score statistic:

$$U \dot{\sim} \mathrm{N}(0, J),$$

where we recall that J is the information

- The above is for a single parameter; in the multiparameter setting,

$$\mathbf{u} \dot{\sim} \mathrm{N}(\mathbf{0}, \mathbf{J}),$$

where $\mathbf{J} = \mathrm{Var}(\mathbf{u})$ is the *information matrix*

- The above approximations are derived from the asymptotic relationship

$$\frac{1}{\sqrt{n}}\mathbf{u} \xrightarrow{\mathsf{d}} \mathrm{N}(\mathbf{0}, \mathbf{J}_1),$$

where $\mathbf{J}_1$ is the expected information from a single observation

The information matrix

- For a single observation, recall that

$$\mathrm{Var}(U_1) = -\mathrm{E}(U_1')$$

- Thus, in multiple dimensions,

$$\mathrm{Var}(\mathbf{u}_1) = -\mathrm{E}\left(\frac{\partial}{\partial\theta}\mathbf{u}_1\right),$$

or in other words,

$$\mathbf{J}_1 = -\mathrm{E}(\mathbf{H}_1),$$

where $\mathbf{H}$ is the Hessian matrix

The information matrix (cont'd)

- If we have multiple observations, and those observations are independent,

$$L(\theta) = \prod_{i=1}^{n} L_i(\theta)$$

$$\ell(\theta) = \sum_{i=1}^{n} \ell_i(\theta)$$

- Furthermore, because derivatives and expectations can be distributed through the summation, we have

$$\mathbf{J} = n\mathbf{J}_1$$

Observed vs. Fisher information

- Furthermore,

$$\begin{aligned}\mathbf{J} &= n\mathbf{J}_1 \\ &= -n\mathrm{E}(\mathbf{H}_1)\end{aligned}$$

- When performing inference (*i.e.*, when we have actual data), it is usually more convenient to simply evaluate the Hessian at the observed data, rather than take the expectation over the data you would have expected to see
- The former is called the *observed information* and the latter the *expected information* or *Fisher information*

Observed vs. Fisher information

- To be more specific:

$$\mathbf{J}(\boldsymbol{\theta}) = -n\mathrm{E}\left\{\mathbf{H}_1(\boldsymbol{\theta}|Y)\right\} \qquad \text{(Fisher)}$$
$$\hat{\mathbf{J}}(\boldsymbol{\theta}) = -\mathbf{H}(\boldsymbol{\theta}|\mathbf{y}) \qquad \text{(Observed)}$$

- Note that in the first equation, Y is a random scalar; in the second, it is a fixed vector
- For the purposes of this class, the distinction between the two is not terribly important (our approximate results hold regardless of which information is used), and I will use $\mathbf{J}$ generically to refer to either kind of information unless otherwise noted

Quadratic approximation to the likelihood

- The preceding results are a useful place to start, but what we really want to know is the sampling distribution of our estimates
- Consider, then, approximating the likelihood as a function of $\boldsymbol{\theta}$
- **Proposition:** The quadratic Taylor series approximation to the likelihood at the MLE is given by

$$\ell(\boldsymbol{\theta}) \approx \ell(\hat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \mathbf{H} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$$

Sampling distribution of MLEs

- Using this approximation, the score is

$$\mathbf{u} \approx \mathbf{H}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}),$$

and we are ready to prove the following result

- **Result:** The sampling distribution of a maximum likelihood estimator is approximately normal, with

$$\hat{\boldsymbol{\theta}} \dot{\sim} \mathrm{N}(\boldsymbol{\theta}, \mathbf{J}^{-1})$$

- More rigorously, it can be shown that under certain regularity conditions,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathsf{d}} \mathrm{N}(\mathbf{0}, \mathbf{J}_1^{-1})$$

Sampling distribution of $\widehat{\boldsymbol{\beta}}$

- The regression coefficients from a GLM are MLEs; it is thus straightforward to show the following
- **Result:** The sampling distribution of the regression coefficients from a GLM are approximately normal, with

$$\widehat{\boldsymbol{\beta}} \dot{\sim} \mathrm{N}\left(\boldsymbol{\beta}, \phi(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\right)$$

- The usual caveat applies: the above is based on the assumption that the model holds

Confidence intervals and hypothesis tests

- We are now in a position to derive confidence intervals and hypothesis tests in manner entirely analogous to our earlier derivations for the linear regression case
- **Result:** Suppose that the model specified by the GLM holds. Then

$$\frac{\hat{\beta}_j - \beta_j}{\widehat{\mathrm{SE}}} \dot{\sim} z,$$

where $\widehat{\mathrm{SE}}$ is the square root of $\hat{\phi}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}_{jj}$

- **Corollary:** Suppose that the model specified by the GLM holds. Then

$$\frac{\boldsymbol{\lambda}^T\widehat{\boldsymbol{\beta}} - \boldsymbol{\lambda}^T\boldsymbol{\beta}}{\widehat{\mathrm{SE}}} \dot{\sim} z,$$

where $\widehat{\mathrm{SE}}$ is the square root of $\hat{\phi}\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\boldsymbol{\lambda}$

Confidence intervals and hypothesis tests (cont'd)

Two details deserve some special attention:

- We're assuming that there is some reasonable way to estimate ϕ; the details vary depending on the distribution
- The matrix of weights, $\mathbf{W}$, is evaluated at the MLE

Wald approach

- In linear regression, we had exact results, and there was a clear, unassailable way of conducting hypothesis tests and constructing confidence intervals
- Not so for generalized linear models: our results are approximate, and there is more than one way to handle the approximation
- The preceding approach, based on taking the asymptotic normality of the MLE literally, is called the *Wald* approach, after Abraham Wald
- The resulting procedures are called "Wald confidence intervals", "Wald hypothesis tests", "Wald test statistics", etc.

Shortcoming of the Wald approach

- The Wald approach has the advantage of simplicity: all you need to know is an estimate and its standard error, and you can construct by hand everything you want to know
- However, the Wald approach depends entirely on the quadratic approximation to the likelihood
- If this approximation is poor, the Wald approach will suffer

The likelihood ratio

- A competing approach is based on likelihood ratios
- Consider two models: one which depends on a vector of parameters $\boldsymbol{\theta}$ and the other which restricts some subset of those parameters to have a known value (we refer to the former as the "full" model and the latter as the "reduced" model)
- Specifically,

$$\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}) \qquad \text{(Full)}$$

$$\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}_0^{(2)}) \qquad \text{(Reduced)},$$

where $\boldsymbol{\theta}_0^{(2)}$ is a specified vector of constants

The likelihood ratio (cont'd)

- The likelihood ratio approach, as the name suggests, is based on the likelihood ratio

$$\lambda = \frac{L_{\text{Full}}}{L_{\text{Reduced}}},$$

or equivalently,

$$\log(\lambda) = \ell_{\text{Full}} - \ell_{\text{Reduced}}$$

- A standard result of likelihood theory is that

$$2\log(\lambda) \dot{\sim} \chi^2_q,$$

where q is the length of $\boldsymbol{\theta}^{(2)}$ (typically, the number of parameters assumed to be zero)

Likelihood ratio tests and confidence intervals

- This result allows us to carry out hypothesis tests by calculating $p = \Pr(\chi^2_q \geq 2\log(\lambda))$
- It also allows us to construct $(1-\alpha)$ confidence intervals by inverting the above test – *i.e.*, finding the set of parameter values $\boldsymbol{\theta}_0^{(2)}$ such that

$$2\left\{\ell(\hat{\boldsymbol{\theta}}) - \ell\left(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}^{(2)} = \boldsymbol{\theta}_0^{(2)}\right)\right\} \leq \chi^2_{\alpha,q},$$

where $\Pr(\chi^2_q \geq \chi^2_{\alpha,q}) = \alpha$

Likelihood ratio vs. Wald

- The Wald approach enjoys popularity due to its simplicity (likelihood ratio confidence intervals are obviously difficult to construct by hand)
- The two approaches often agree quite well
- However, there are also situations where the two disagree dramatically
- Tests and confidence intervals based on likelihood ratios are more accurate, and should always be trusted over the Wald approach