Inference: Likelihood ratio vs. Wald approaches

Patrick Breheny

March 19
Thus far, all our inferences have been based on the result:

\[ \hat{\beta} \sim N(\beta, \phi(X^T W X)^{-1}) \]

This approach has the great advantage of simplicity: all you need to know is \( \hat{\beta} \) and \( \text{Var}(\hat{\beta}) \) and you may construct by hand all the tests and confidence intervals you need for any element of \( \beta \) or any linear combination of the elements of \( \beta \) (these are called “Wald tests”, “Wald confidence intervals”, etc.)

Recall, however, that the result on the previous slide is based on an approximation to the likelihood at the MLE, and this approximation may be poor at \( \beta \) values far from \( \hat{\beta} \)
A competing approach is based on likelihood ratios

We consider the univariate case first, comparing the likelihood at an arbitrate value $\theta$ with that of the MLE $\hat{\theta}$:

$$\lambda = \frac{L(\theta)}{L(\hat{\theta})}$$

**Theorem:** As $n \to \infty$ with iid data, subject to the usual regularity conditions,

$$-2 \log \lambda \xrightarrow{d} \chi^2_1$$
This result can be extended to multivariate and non-iid cases as well; consider two models:

Full: $\mathbf{\beta} = (\beta^{(1)}, \beta^{(2)})$

Reduced: $\mathbf{\beta} = (\beta_0^{(1)}, \beta^{(2)})$

where $\beta_0^{(1)}$ is a specified vector of constants

Letting $\lambda$ denote the likelihood ratio comparing the reduced model to the full model, we have

$$-2 \log \lambda \sim \chi^2_q,$$

where $q$ is the length of $\beta^{(1)}$ (typically, the number of parameters assumed to be zero)
This result allows us to carry out hypothesis tests by calculating \( p = \Pr(\chi_q^2 \geq 2 \log(\lambda)) \).

It also allows us to construct \((1 - \alpha)\) confidence intervals by inverting the above test – i.e., finding the set of parameter values \( \beta_0^{(1)} \) such that

\[
-2 \log \frac{L(\hat{\beta}|\beta^{(1)} = \beta_0^{(1)})}{L(\hat{\beta})} \leq \chi_{1-\alpha,q}^2,
\]

where \( \chi_{1-\alpha,q}^2 \) is the \((1 - \alpha)\) quantile of the \( \chi^2 \) distribution with \( q \) degrees of freedom.
Wald vs. Likelihood ratio

Estimating the effect of age upon survival for females in the Donner party:

\[ 2l(\beta) - l(\hat{\beta}) \]

95% confidence intervals:

Wald: \((-0.365, -0.023)\)
LR: \((-0.428, -0.057)\)
Remarks

- As you can see, the Wald approach is incapable of capturing asymmetry in the likelihood function, and must therefore always produce symmetric confidence intervals about the MLE.
- The likelihood ratio is not subject to this restriction (the downside, of course, is that we must refit a new model at all the different values for $\beta$).
- This impacts hypothesis testing as well: for testing the interaction term, the Wald test gives $p = 0.087$ while the LRT gives $p = 0.048$. 
Wald vs. Likelihood ratio

For the donner data, \( n = 45 \) and \( p = 3 \); when \( n \) is larger, the agreement is much better (here, \( n = 100, \ p = 2 \)):

\[
2 \{ l(\hat{\beta}) - l(\beta) \}
\]

95% confidence intervals:

Wald: \((-0.321, 0.461)\)
LR: \((-0.322, 0.468)\)
When $n$ is smaller, the agreement is even worse (here, $n = 6$, $p = 2$):

![Graph showing likelihood ratio and Wald tests](image)

95% confidence intervals:

Wald: $(-10.4, 35.3)$
LR: $(0.336, 59.7)$
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 be trusted over the Wald approach
Complete separation

- Just as in univariate statistics, when $n$ is large we can often ignore the fact that our data is discrete and use a normal approximation.
- When $n$ is small, however, problems can arise.
- Consider the following data:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>$y$</td>
</tr>
<tr>
<td>-1.64</td>
<td>0</td>
</tr>
<tr>
<td>-0.80</td>
<td>0</td>
</tr>
<tr>
<td>-0.46</td>
<td>0</td>
</tr>
<tr>
<td>-0.46</td>
<td>0</td>
</tr>
<tr>
<td>-0.34</td>
<td>0</td>
</tr>
<tr>
<td>0.12</td>
<td>1</td>
</tr>
<tr>
<td>0.62</td>
<td>1</td>
</tr>
<tr>
<td>0.64</td>
<td>1</td>
</tr>
<tr>
<td>0.73</td>
<td>1</td>
</tr>
<tr>
<td>1.10</td>
<td>1</td>
</tr>
</tbody>
</table>
If we try to fit a logistic regression model to this data, we find that the algorithm will not converge and we get warning messages in SAS and R.

The reason is that all of the events occur when $x$ is large and don’t occur when $x$ is small.

To put it another way, we can draw a line in the $x$’s and separate the $y = 0$’s from the $y = 1$’s.

This phenomenon is referred to as complete separation (or more generally, as the problem of monotone likelihood).
Monotone likelihood
Ramifications

- What it means is that the MLE occurs at infinity (or $-\infty$)
- This has a number of ramifications:
  - Numerical algorithms will fail
  - Weights will go to zero
  - Standard errors will go to infinity
- Note, however, that likelihood ratio tests are still valid
Complete separation: Practical aspects

- This has a number of complicated ramifications for inference lie beyond the scope of this course.
- Practically speaking, the ramifications are that the data do not allow you to estimate a certain parameter in the way that the model is currently specified.
- This can often occur when models are overparameterized – in models with many explanatory variables, complete separation occurs whenever a linear predictor completely separates the outcome.
- In linear regression, estimates are only undefined if $X$ is not full rank; in logistic regression, complete separation represents an additional restriction on the complexity of the design matrix.
Fitted probabilities of 0 or 1

Finally, it is worth noting that you may sometimes see a warning message along the lines of “fitted probabilities numerically 0 or 1 occurred”; this is very different from complete separation.

Because $\pi_i$ is a function of $\exp(\eta_i)$, extreme $\eta_i$ values can easily produce fitted probabilities extremely close to 0 or 1; this causes problems numerically in the IRLS algorithm, since $W_i = \pi_i(1 - \pi_i)$.

Keep in mind that this is a warning, not an error – the model can still be fit and all the usual inferential procedures applied.

However, it is generally an indication that your data contains outliers, and some investigation into those points with 0 or 1 probabilities is typically warranted.