Sampling distribution of GLM regression coefficients

Patrick Breheny

February 5
Introduction

- So far, we’ve discussed the basic properties of the score, and the special connection between the score and the natural parameter ($\theta$) that exists in exponential families.
- Today, in the final installment of our three-part series on likelihood theory, we’ll arrive at the important result: what does all this imply about the distribution of the maximum likelihood estimator, $\hat{\theta}$?
The basic mathematical tool we will need for today is the Taylor series expansion, one of the most widely applicable and useful tools in statistics. The basic idea is to take a complicated function and simplify it by approximating it with a straight line:

\[ f(x) \approx f(x_0) + f'(x_0)(x - x_0), \]

where \( x_0 \) is the point we are basing the approximation on. This approximation will be reasonably accurate provided that we are in the neighborhood of \( x_0 \).
Taylor series expansions: Illustration
The idea can be extended to higher-order polynomials as well:

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

provides a quadratic approximation to \( f(x) \)

This will provide an even more accurate approximation

In principle, one could keep going with higher and higher order derivatives, obtaining more and more accurate approximations, but all we need for the purposes of this class is first- and second-order approximations
The preceding formulas are for univariate functions; the idea readily extends to functions of more than one variable:

\[
f(x) \approx f(x_0) + \nabla f(x_0)^T(x - x_0)
\]

\[
f(x) \approx f(x_0) + (x - x_0)^T \nabla f(x_0) + \frac{1}{2}(x - x_0)^T \{\nabla^2 f(x_0)\}(x - x_0)
\]

This are the versions we need for regression modeling, as we have quantities (e.g., the likelihood, the score) that will depend on a vector of parameters \( \beta \)
Recall that

\[ u \sim N(0, V) \]

and that, for exponential families,

\[ u = \sum_{i=1}^{n} \frac{Y_i - b'(\theta)}{\phi} \]

Thus, we know the (approximate) distribution of \( u \), but the distribution of \( \hat{\theta} \) is complex because the function \( b'(\theta) \) may be complicated and nonlinear.
Approximating the score

- We can make progress, however, by applying a Taylor series approximation to the score at the MLE.
- Let $H(\theta) = \nabla u(\theta)$; note that this is the Hessian matrix of second derivatives for the log-likelihood.
- **Result:**

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

  where $\hat{\theta}$ is the MLE; or more simply,

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

  provided we keep in mind that $H$ is evaluated at the MLE.
Observed vs. Fisher information

- Recall that there was a connection between the Hessian and the information:

\[ V = -E(H); \]

in other words, the information is the (negative) Hessian we would expect to observe

- In practice, it usually easier to deal with \( H(\hat{\theta}) \), the Hessian we actually did observe

- Correspondingly, \( -H(\theta|y) \) is referred to as the *observed information*, as opposed to \( -nE\{H(\theta|Y)\} \), which is referred to as the *Fisher information*
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.

I will use the term “information” and the symbol $\mathbf{V}$ generically to refer to either kind of information, unless otherwise noted.
We are now ready to prove the following:

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

\[ \hat{\theta} \sim N(\theta, V^{-1}) \]

This can also be stated more rigorously; under certain regularity conditions,

\[ \sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V_{i}^{-1}) \]
Remarks

- Note that this relationship provides another perspective on information: as the information in the sample goes up, the variability of $\hat{\theta}$ goes down (as does, correspondingly, our uncertainty about the true value of $\theta$).
- This also allows us to use familiar results from the normal distribution to construct tests and confidence intervals for individual parameters $\theta_j$.
- Furthermore, it tells us how the MLEs for various parameters are correlated, allowing us to easily work out the sampling distributions for linear combinations of parameters.
Another way of thinking about what we are doing is as a quadratic approximation to the log-likelihood:

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

Noting that the log-likelihood of the normal distribution actually is quadratic; it should come as no surprise that $\hat{\theta}$ is normally distributed.
The preceding derivations have all assumed we have identically and independently distributed observations.

This, of course, is not the case in modeling: the natural parameter, $\theta_i$, for each observation is different; we expect it to change depending on the explanatory variables – indeed, understanding how the explanatory variables affect the outcome is the entire point of the analysis.

Of course, we don’t go about estimating $\{\theta_i\}$ directly, as this would be unstable; instead, we impose a relationship between $\theta$ and the explanatory variables that is governed by the systematic component of the model.
In what follows, I will assume we are working with the canonical link, in which case we are directly modeling the natural parameters and $\theta = \mathbf{X}\beta$; you can still work out relationships and distributions for other links, but the expressions are quite a bit messier.

Specifically, for the canonical link, $\frac{\partial \theta}{\partial \beta} = \mathbf{X}^T$; this greatly simplifies the application of the chain rule in what follows.
Provided that we are estimating $\beta$ using maximum likelihood, we can apply our earlier result and state that

$$\hat{\beta} \sim N(\beta, V^{-1});$$

the only catch is that we have to work out the information with respect to $\beta$.

**Theorem:** For the canonical link,

$$V = \phi^{-1}X^T W X,$$

where $W$ is an $n \times n$ diagonal matrix with entries $W_{ii} = W(\mu_i)$, the function dictating the mean-variance relationship for distribution in an exponential family.
To summarize, then, we have the following theorem:

**Theorem:** The sampling distribution of the regression coefficients from a GLM with canonical link are approximately normal, with

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

The usual caveat applies: the above is based on the assumption that the model holds.
Confidence intervals and hypothesis tests

- Thus, we can derive confidence intervals and hypothesis tests in manner entirely analogous to the linear regression case.

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

  \[
  \frac{\hat{\beta}_j - \beta_j}{\hat{SE}} \sim Z,
  \]

  where \(\hat{SE}\) is the square root of \(\hat{\phi}(X^TWX)^{-1}_{jj}\)

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

  \[
  \frac{\lambda^T\hat{\beta} - \lambda^T\beta}{\hat{SE}} \sim Z,
  \]

  where \(\hat{SE}\) is the square root of \(\hat{\phi}\lambda^T(X^TWX)^{-1}\lambda\)
Note that:

- We’re assuming that there is some reasonable way to estimate \( \phi \); the details vary depending on the distribution
- The matrix \( \mathbf{W} \) is evaluated at \( \hat{\beta} \)

Furthermore, recall that this is an approximation based on the MLE; as we saw at the beginning, this approximation may not be accurate for \( \beta \) far away from \( \hat{\beta} \)

We’ll look at the implications of this, as well as remedies, later in the semester