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 (θ) 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, $\hat{\theta}$?

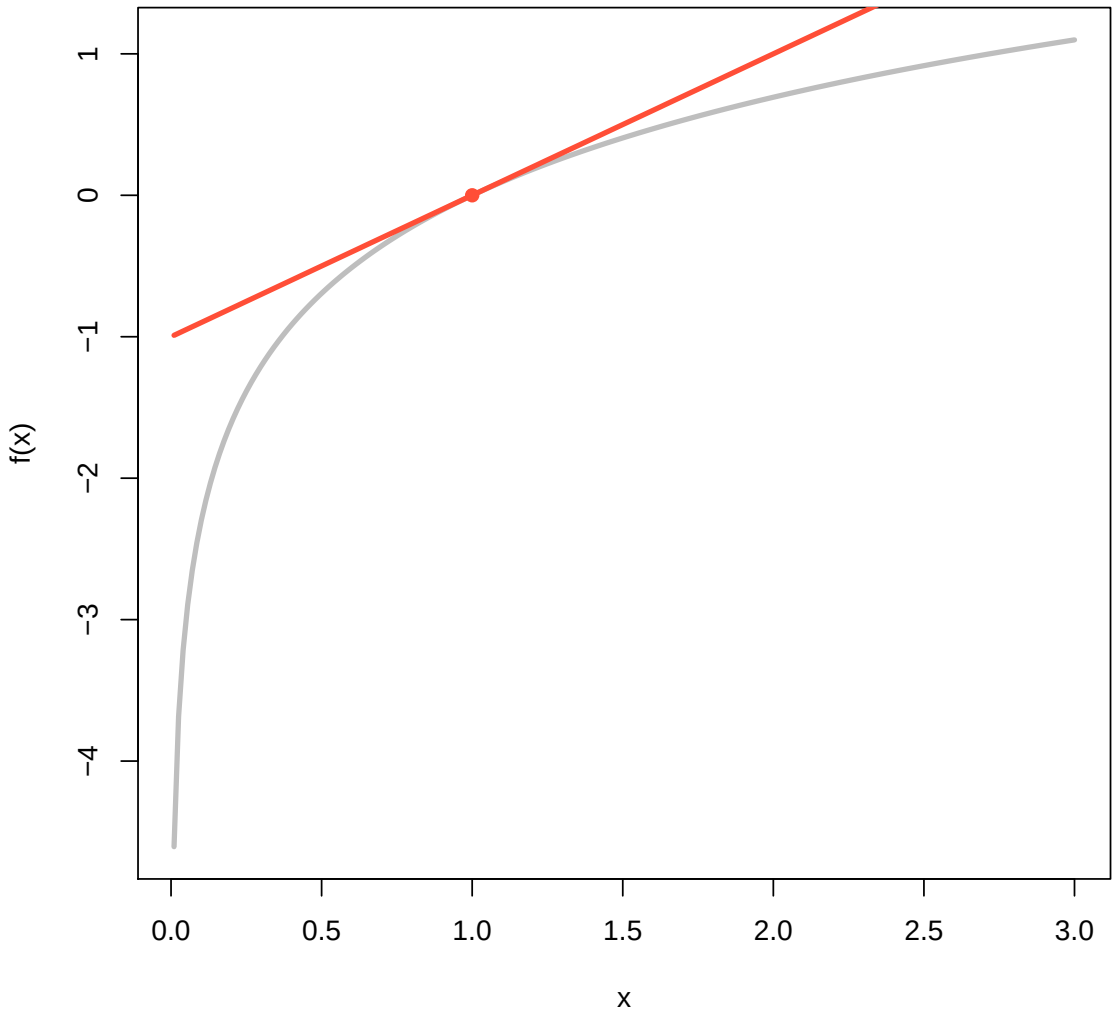
Taylor series expansions

- 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



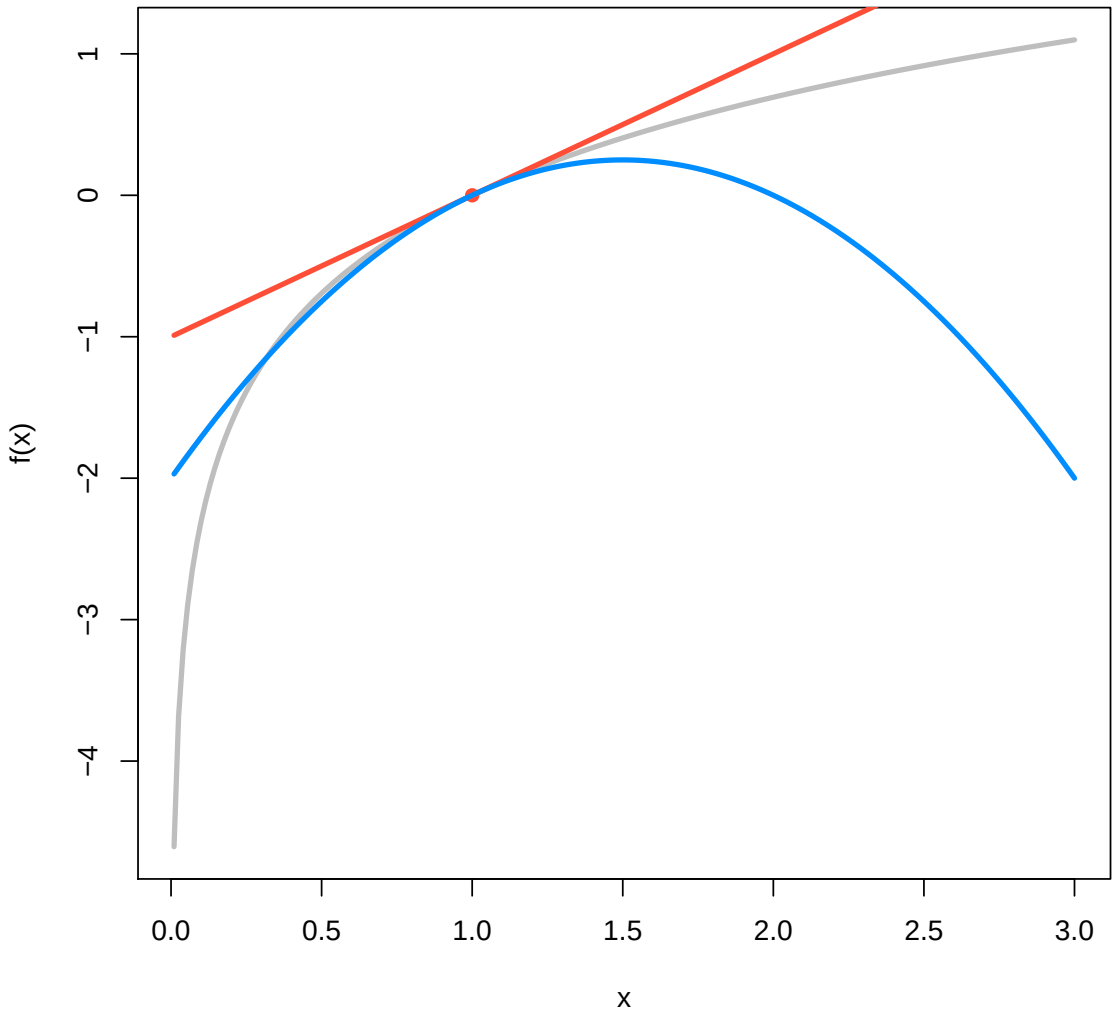
Quadratic approximations

- 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

Quadratic illustration



Multivariate extensions

- The preceding formulas are for univariate functions; the idea readily extends to functions of more than one variable:

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^T(\mathbf{x} - \mathbf{x}_0)$$

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + (\mathbf{x} - \mathbf{x}_0)^T \nabla f(\mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T \{\nabla^2 f(\mathbf{x}_0)\}(\mathbf{x} - \mathbf{x}_0)$$

- 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 $\boldsymbol{\beta}$

Relationship between the score and the MLE

- Recall that

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

and that, for exponential families,

$$\mathbf{u} = \sum_{i=1}^{n} \frac{Y_i - b'(\boldsymbol{\theta})}{\phi}$$

- Thus, we know the (approximate) distribution of $\mathbf{u}$, but the distribution of $\hat{\boldsymbol{\theta}}$ is complex because the function $b'(\boldsymbol{\theta})$ 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 $\mathbf{H}(\boldsymbol{\theta}) = \nabla \mathbf{u}(\boldsymbol{\theta})$; note that this is the Hessian matrix of second derivatives for the log-likelihood
- **Result:**

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

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

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

provided we keep in mind that $\mathbf{H}$ is evaluated at the MLE

Observed vs. Fisher information

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

$$\mathbf{V} = -\mathrm{E}(\mathbf{H});$$

  in other words, the information is the (negative) Hessian we would expect to observe
- In practice, it usually easier to deal with $\mathbf{H}(\hat{\boldsymbol{\theta}})$, the Hessian we actually did observe
- Correspondingly, $-\mathbf{H}(\boldsymbol{\theta}|\mathbf{y})$ is referred to as the *observed information*, as opposed to $-n\mathrm{E}\{\mathbf{H}(\boldsymbol{\theta}|Y)\}$, which is referred to as the *Fisher information*

Observed vs. Fisher information (cont'd)

- 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

Sampling distribution of MLEs

- We are now ready to prove the following:
- **Theorem:** The sampling distribution of a maximum likelihood estimator is approximately normal, with

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

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

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

Remarks

- Note that this relationship provides another perspective on information: as the information in the sample goes up, the variability of $\hat{\boldsymbol{\theta}}$ goes down (as does, correspondingly, our uncertainty about the true value of θ)
- This also allows us to use familiar results from the normal distribution to construct tests and confidence intervals for individual parameters θ_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

Remarks

- Another way of thinking about what we are doing is as a quadratic approximation to the log-likelihood:

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

- Noting that the log-likelihood of the normal distribution actually *is* quadratic; it should come as no surprise that $\hat{\boldsymbol{\theta}}$ is normally distributed

Non-identically distributed observations

- 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, θ_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 $\boldsymbol{\theta}$ and the explanatory variables that is governed by the systematic component of the model

Canonical link simplification

- In what follows, I will assume we are working with the canonical link, in which case we are directly modeling the natural parameters and $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$; you can still work out relationships and distributions for other links, but the expressions are quite a bit messier
- Specifically, for the canonical link, $\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\beta}} = \mathbf{X}^T$; this greatly simplifies the application of the chain rule in what follows

Information under the canonical link

- Provided that we are estimating $\boldsymbol{\beta}$ using maximum likelihood, we can apply our earlier result and state that

$$\widehat{\boldsymbol{\beta}} \dot{\sim} \mathrm{N}(\boldsymbol{\beta}, \mathbf{V}^{-1});$$

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

- **Theorem:** For the canonical link,

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

  where $\mathbf{W}$ is an $n \times n$ diagonal matrix with entries $\mathbf{W}_{ii} = W(\mu_i)$, the function dictating the mean-variance relationship for distribution in an exponential family

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

- 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

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

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

Confidence intervals and hypothesis tests

- 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

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

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)

- Note that:
  - We're assuming that there is some reasonable way to estimate ϕ; the details vary depending on the distribution
  - The matrix $\mathbf{W}$ is evaluated at $\widehat{\boldsymbol{\beta}}$
- Furthermore, recall that this is an approximation based on the MLE; as we saw at the beginning, this approximation may not be accurate for $\boldsymbol{\beta}$ far away from $\widehat{\boldsymbol{\beta}}$
- We'll look at the implications of this, as well as remedies, later in the semester