Score and information

Patrick Breheny

October 2

Introduction

- In our previous lecture, we saw how likelihood-based inference works for exponential families
- Starting today, we are going to adopt a more general outlook on likelihood, and not make any specific assumptions about its form
- As we remarked at the outset of the course, the likelihood function is minimal sufficient
- This means that the *entire function* is the object that contains the information necessary for objective inference

Maximum likelihood estimation

- However, a number is of course much simpler and easier to communicate and manipulate than an entire function, so it is desirable to summarize and simplify the likelihood
- The single most important information about the likelihood is surely the value at which it is maximized
- The *maximum likelihood estimator*, $\hat{\boldsymbol{\theta}}$, of a parameter $\boldsymbol{\theta}$, given observed data $\mathbf{x}$, is

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathbf{x}).$$

- This was Fisher's original motivation for the likelihood, as a means of estimating scientific quantities of interest (in his later years, however, he no longer thought of likelihood as merely a device for producing point estimates)

Curvature

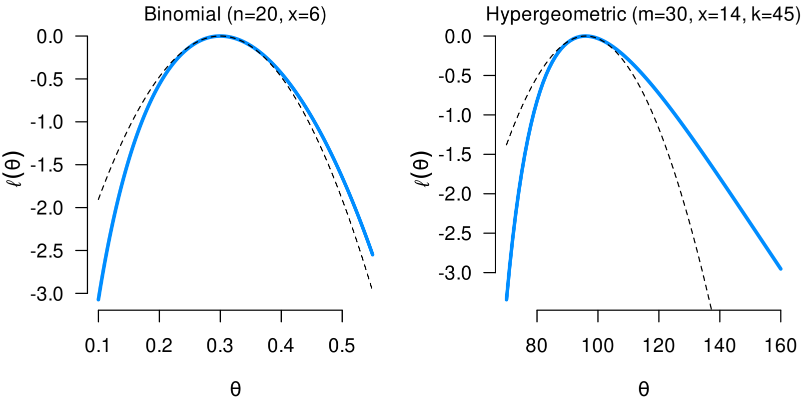
- A single number is not enough to represent a function
- However, if the likelihood function is approximately quadratic, then two numbers are enough to represent it: the location of its maximum and its curvature at the maximum
- Specifically, what I mean by this is that any quadratic function can be written

$$f(x) = c(x - m)^2 + \text{Const},$$

where c is the curvature and m the location of its maximum; the constant is irrelevant given our earlier remarks about how only likelihood comparisons are only meaningful in the relative sense

Quadratic approximation: Illustration

The likelihood itself does not tend to be quadratic, but the *log-likelihood* does; revisiting the two examples from our first lecture:



Remarks

- Log is a monotone function, so the value of θ that maximizes the log-likelihood also maximizes the likelihood
- Even good approximations break down for $\boldsymbol{\theta}$ far from $\hat{\boldsymbol{\theta}}$: regularity is a local phenomenon
- As we will be referring to it often, we will use the symbol ℓ to denote the log-likelihood: $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$
- The situation is similar in multiple dimensions; any quadratic function can be written

$$f(\mathbf{x}) = (\mathbf{x} - \mathbf{m})^\top \mathbf{C}(\mathbf{x} - \mathbf{m}) + \text{Const};$$

  we now require a $d \times 1$ vector $\mathbf{m}$ to denote the location of the maximum and a $d \times d$ matrix $\mathbf{C}$ to describe the curvature

Regularity

- Likelihood functions that can be adequately represented by a quadratic approximation are called *regular*[1]
- Conditions that ensure the validity of the approximation are called *regularity conditions*
- We will discuss regularity conditions in detail later; for now, we will just assume that the likelihood is regular

[1] When we say that the likelihood has a quadratic approximation, what we really mean of course is that the log-likelihood has a quadratic approximation

The score statistic

- The derivative of the log-likelihood is a critical quantity for describing this quadratic approximation
- The quantity is so important that it is given its own name in statistics, the *score*, and often denoted $\mathbf{u}$:

$$\mathbf{u}(\boldsymbol{\theta}) = \nabla \ell(\boldsymbol{\theta}|\mathbf{x})$$

- Note that
  - $\mathbf{u}$ is a function of θ
  - For any given $\boldsymbol{\theta}$, $\mathbf{u}(\boldsymbol{\theta})$ is a random variable, as it depends on the data $\mathbf{x}$; usually suppressed in notation
  - For independent observations, the score of the entire sample is the sum of the scores for the individual observations:

$$\mathbf{u}(\boldsymbol{\theta}) = \sum_i \mathbf{u}_i(\boldsymbol{\theta})$$

Score equations

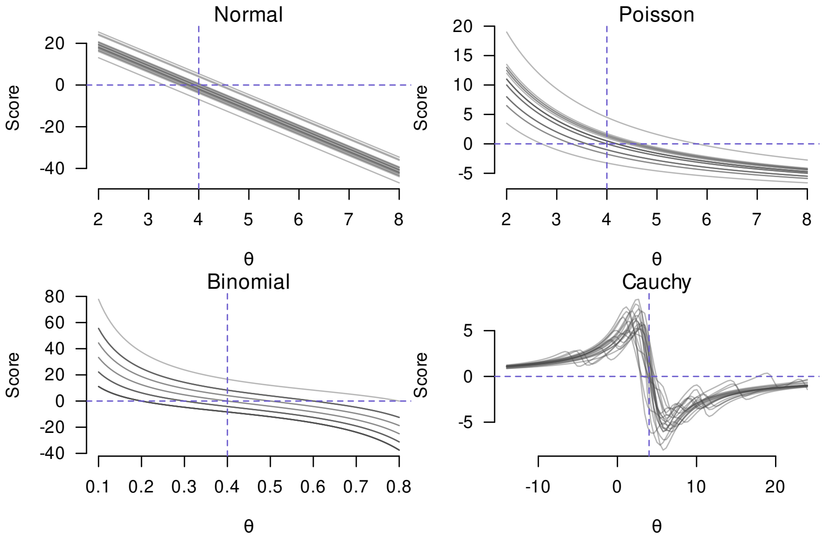
- If the likelihood is regular, we can find $\hat{\boldsymbol{\theta}}$ by setting the gradient equal to zero; the MLE is the solution to the equation(s)

$$\mathbf{u}(\boldsymbol{\theta}) = \mathbf{0};$$

  this system of equations is known as the *score equation(s)* or sometimes the *likelihood equation(s)*

- For example, suppose we have $X_i \overset{\text{iid}}{\sim} \text{N}(\theta, \sigma^2)$ with σ^2 known
  - $U_i(\theta) = (X_i - \theta)/\sigma^2$
  - $U(\theta) = \sum_i (X_i - \theta)/\sigma^2$
  - $U(\hat{\theta}) = 0 \implies \hat{\theta} = \bar{x}$

Illustration (vertical line at θ^*)



Information

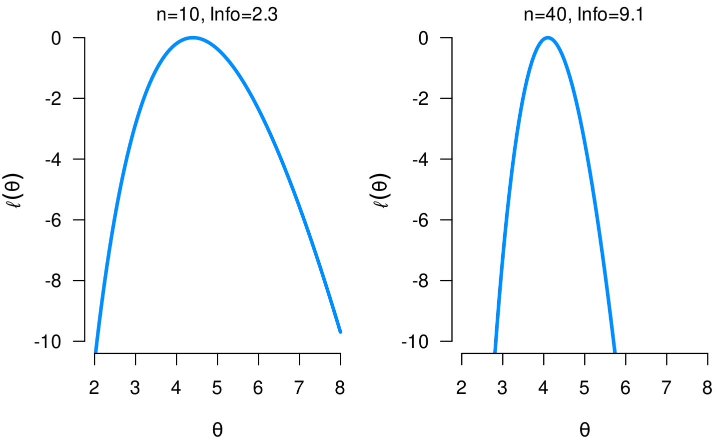
- Meanwhile, the curvature is given by the second derivative
- This quantity is called the *information*,

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = -\nabla^2 \ell(\boldsymbol{\theta});$$

  the negative sign arises because the curvature at the maximum is negative
- The name "information" is an apt description: the larger the curvature, the sharper (less flat) the peak, so the less uncertainty we have about $\boldsymbol{\theta}$

Information: Illustration

Random sample from the Poisson distribution:



Information: Example

- As an analytic example, let's return to the situation with $X_i \overset{\text{iid}}{\sim} \text{N}(\theta, \sigma^2)$ and σ^2 known
  - $\mathcal{I}_i(\theta) = 1/\sigma^2$
  - $\mathcal{I}(\theta) = n/\sigma^2$
- Note that
  - For independent samples, the total information is the sum of the information obtained from each observation
  - Noisier data $\implies$ less information
- In general, the information depends on both X and θ (the normal is a special case); we'll return to this point later

Information: Another example

- As another example, suppose there are 5 observations taken from a $\mathrm{N}(\theta, 1)$ distribution, but we observe only the maximum $x_{(5)} = 3.5$
- Here, it is not clear how we would find the MLE, score, and information analytically, but we can use numerical procedures to optimize and calculate derivatives
- In this case, the information is 2.4, implying that knowing the maximum of 5 observations is worth 2.4 observations – better than a single observation, but not as good as having all 5 observations

Normal likelihood

- From an inferential standpoint, we can view this quadratic approximation as a normal approximation, as a quadratic log-likelihood corresponds to the Gaussian distribution
- As we mentioned in our first class, connecting likelihood to probability is challenging in general; however, it is easy in the case of the normal distribution
- For an iid sample from a $\mathrm{N}(\theta, \sigma^2)$ distribution (assuming σ^2 known; we'll consider the multiparameter case next), the likelihood is

$$
\begin{aligned}
L(\theta) &\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_i (x_i - \theta)^2\right\} \\
&\propto \exp\left\{-\frac{n}{2\sigma^2}(\bar{x} - \theta)^2\right\}
\end{aligned}
$$

Likelihood ratios

- The likelihood ratio, then, is simply

$$\log \frac{L(\theta)}{L(\hat{\theta})} = -\frac{n}{2\sigma^2}(\bar{x} - \theta)^2$$

- Furthermore, letting θ^* denote the true value of θ, we know that $(\bar{x} - \theta^*)/(\sigma/\sqrt{n}) \sim \mathrm{N}(0, 1)$, so

$$2 \log \frac{L(\hat{\theta})}{L(\theta^*)} \sim \chi^2_1$$

- This means that the likelihood interval $\{\theta : L(\theta)/L(\hat{\theta}) \geq c\}$ has (frequentist) probability $\mathbb{P}(\chi^2_1 \leq -2 \log c)$ of containing θ^*

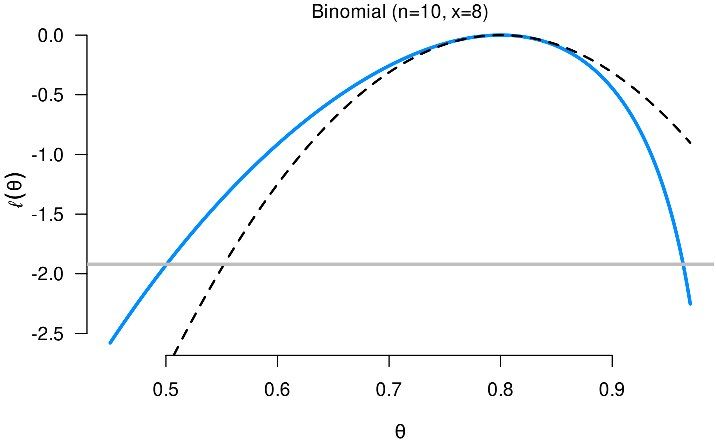
Standard errors

- In other words, if we want a 95% confidence interval, we should set $c = \exp\{-\frac{1}{2}\chi^2_{1,(.95)}\} \approx 0.15$
- Furthermore, solving for the endpoints of the interval, we have

$$\bar{x} \pm \sqrt{-2\log c} \cdot (n/\sigma^2)^{-1/2},$$

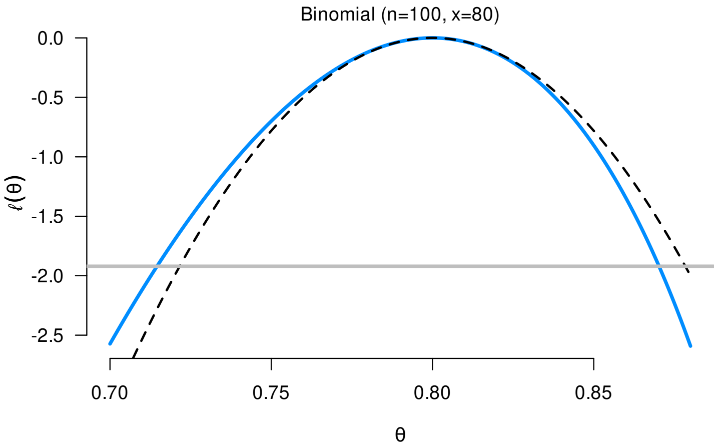
  or $\hat{\theta} \pm z\mathrm{SE}$, where $\mathrm{SE} = (n/\sigma^2)^{-1/2} = \mathcal{I}^{-1/2}$ and z is an appropriate quantile of the standard normal distribution
- These probabilities are exact in the special case of the normal distribution with known variance, but it stands to reason that they should be approximately correct if the likelihood is regular (we'll formalize this idea in the coming lectures)

Binomial illustration (n=10, $\theta = 0.8$)



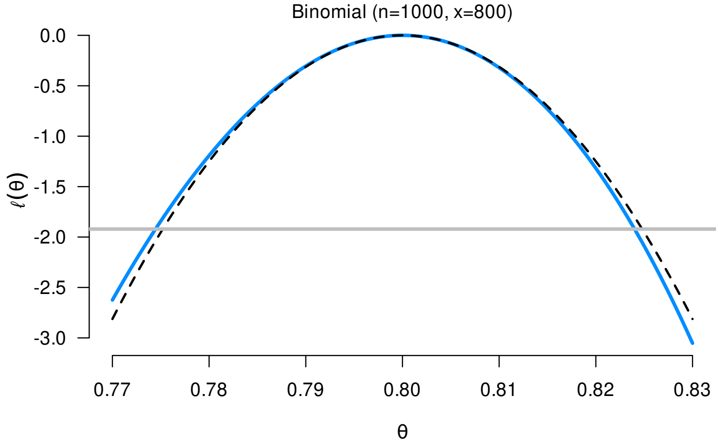
Actual coverage (simulation): 88.3%

Binomial illustration (n=100, $\theta = 0.8$)



Actual coverage (simulation): 93.2%

Binomial illustration (n=1000, $\theta = 0.8$)



Actual coverage (simulation): 94.9%

Multiparameter case

- Similarly, for the multivariate normal (assuming a nonsingular variance),

$$\log \frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})} = -\tfrac{1}{2}(\bar{\mathbf{x}} - \boldsymbol{\theta})^\top \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\theta}),$$

  so the likelihood interval $\{\boldsymbol{\theta} : L(\boldsymbol{\theta})/L(\hat{\boldsymbol{\theta}}) \geq c\}$ has probability $\mathbb{P}(\chi^2_d \leq -2\log c)$ of containing $\boldsymbol{\theta}^*$
- Note that the presence of multiple parameters changes the probability calibration; for example, with $d = 5$
  - $c = 0.15$ now provides only a 0.42 probability of containing θ^*
  - We now need $c = 0.004$ to attain 95% coverage

"Pure" likelihood for multiparameter problems?

- The interval $\{\theta : L(\theta)/L(\hat{\theta}) \geq c\}$ is based purely on likelihood; as we remarked in our first lecture, the interval itself is neither Bayesian nor frequentist – those paradigms arise only in attempting to assign this interval a probability
- Is a "pure" likelihood approach possible in the multiparameter case (i.e., without the frequentist χ^2 calculations to guide us)?
- Suppose the (relative) likelihood of each parameter is (approximately) independent so that, for example, if $L(\theta_1) = 0.2$ and $L(\theta_2) = 0.2$, then $L(\boldsymbol{\theta}) = 0.2^2 = 0.04$
- Using $c = 0.15$ leads to something of a contradiction: θ_1 and θ_2 are both "likely", but somehow the pair (θ_1, θ_2) is "unlikely"

"Pure" likelihood for the multiparameter case

- An obvious solution is to use c^d: now if $L(\boldsymbol{\theta}) < 0.15^2$, then we must have $L(\theta_1) < 0.15$ or $L(\theta_2) < 0.15$
- Furthermore, we can write $\{\theta : L(\boldsymbol{\theta})/L(\hat{\boldsymbol{\theta}}) \geq c^d\}$ as

$$2\ell(\boldsymbol{\theta}) - 2\ell(\hat{\boldsymbol{\theta}}) \geq 2d \log c,$$

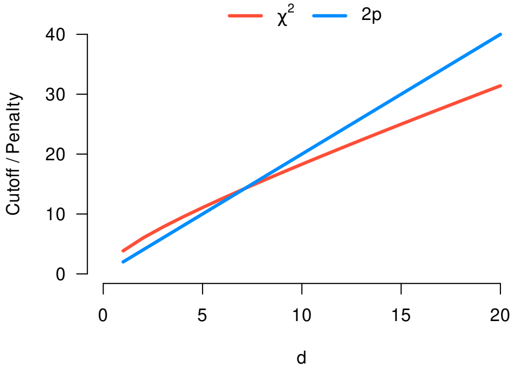
or, using the specific value $c = e^{-1}$,

$$-2\ell(\hat{\boldsymbol{\theta}}) + 2d \geq -2\ell(\boldsymbol{\theta})$$

- In other words, we have arrived at the AIC criterion: $\hat{\boldsymbol{\theta}}$ is an attractive model, despite adding d parameters, if the above inequality holds
- Note that $c = e^{-1} = 0.37$, quite a bit larger than the $c = 0.15$ the likelihood ratio test would imply

AIC vs χ^2

However, this relationship changes as d grows: from a model selection perspective, AIC is more permissive at low d, but more restrictive at larger d



Properties of the score: Introduction

- Earlier, we defined the score as the random function $\mathbf{u}(\boldsymbol{\theta}) = \nabla \ell(\boldsymbol{\theta}|\mathbf{x})$
- With some mild conditions, the random variable $\mathbf{u}(\boldsymbol{\theta}^*)$ turns out to have some rather elegant properties
- These properties are at the core of proving many important results about likelihood theory

Expectation

- We saw earlier that $\mathbf{u}(\boldsymbol{\theta}^*)$ tends to vary randomly about zero; let us now formalize this observation
- **Theorem:** Suppose the likelihood allows its gradient to be passed under the integral sign. Then $\mathbb{E}\mathbf{u}(\boldsymbol{\theta}^*) = \mathbf{0}$.
- A derivative is a type of limit, so whether or not it can be passed under the integral sign is governed by the dominated convergence theorem
- In this particular context, note that $\boldsymbol{\theta}^*$ cannot be on the boundary of the parameter space, and that $\|\nabla L(\boldsymbol{\theta}|x)\| \leq g(x)$, where $\mathbb{E}g(X) < \infty$

Variance of the score

- Under similar conditions involving the second derivative, we also have a nice result involving the variance: namely, that the variance of the score is the expected information
- The variance of the score is called the *Fisher information*, which we will denote $\mathscr{I}$: $\mathscr{I}(\boldsymbol{\theta}) = \mathbb{V}\mathbf{u}(\boldsymbol{\theta}|X)$; its connection with our previous definition of information is made clear in the following theorem
- **Theorem:** Suppose the likelihood allows its Hessian to be passed under the integral sign. Then $\mathscr{I}(\boldsymbol{\theta}^*) = \mathbb{E}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*|X)$.
- This requires the same sort of smoothness conditions as before, except now applied to the second derivatives

Remarks

- Recall that the information $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = -\nabla^2 \ell(\boldsymbol{\theta})$ depends on the data X
- By taking an expected value, we are essentially averaging over different data sets that could occur, weighted by their probability
- To distinguish between the two, the information using the observed data is called the *observed information*
- Note: Keep in mind that that $\boldsymbol{\mathcal{I}}$ is random, while $\boldsymbol{\mathscr{I}}$ is fixed

Notation

- We have something of a notation dilemma, as we need to distinguish between the "total" information from all observations and the information you would expect to get from a single sample, as well as between the observed and Fisher information
- In this class, I'll use the convention of using $\mathcal{I}$ to denote the total observed information, while $\mathscr{I}$ denotes the expected information per observation; i.e., $\mathcal{I} \approx n\mathscr{I}$ (we expect)

Distribution

- Furthermore, since $\mathbf{u}(\boldsymbol{\theta}|\mathbf{x}) = \sum_i \mathbf{u}(\boldsymbol{\theta}|x_i)$, we can apply the central limit theorem to see that

$$\sqrt{n}\{\bar{\mathbf{u}}(\boldsymbol{\theta}^*) - \mathbb{E}\mathbf{u}(\boldsymbol{\theta}^*)\} \xrightarrow{\mathrm{d}} \mathrm{N}(\mathbf{0}, \boldsymbol{\mathscr{I}}(\boldsymbol{\theta}^*)),$$

or

$$\frac{\mathbf{u}(\boldsymbol{\theta}^*)}{\sqrt{n}} \xrightarrow{\mathrm{d}} \mathrm{N}(\mathbf{0}, \boldsymbol{\mathscr{I}}(\boldsymbol{\theta}^*))$$

- Showing that the maximum likelihood estimators, on the other hand, are asymptotically normal (thereby justifying our earlier normal-based inferential procedures) involves a bit more work (we'll take up this question in a later lecture)

Observed vs expected information

- Earlier, we discussed the idea that the width of, say, confidence intervals depends on the information
- We've now introduced two kinds of information; which should we use for inferential purposes?
- Broadly speaking, either one is fine: by the WLLN, $\frac{1}{n}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) \xrightarrow{\text{P}} \boldsymbol{\mathscr{I}}(\boldsymbol{\theta})$, so we have both

$$\frac{1}{\sqrt{n}}\boldsymbol{\mathscr{I}}(\boldsymbol{\theta}^*)^{-1/2}\mathbf{u}(\boldsymbol{\theta}^*) \xrightarrow{\text{d}} \text{N}(\mathbf{0}, \mathbf{I})$$

and

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*)^{-1/2}\mathbf{u}(\boldsymbol{\theta}^*) \xrightarrow{\text{d}} \text{N}(\mathbf{0}, \mathbf{I})$$

assuming $\boldsymbol{\mathscr{I}}$ and $\boldsymbol{\mathcal{I}}$ are positive definite

Observed vs expected information (cont'd)

- At the same time, with any finite sample, they aren't the same . . . surely one tends to be better than the other?
- We'll come back to this point later in the course, but yes, observed information does tend to be better than expected information
- Often, however, practical considerations outweigh theoretical ones
- When the Fisher information is easy to calculate and results in cancellation/simplification, it is often used

Observed vs expected information (cont'd)

- Conversely, sometimes the Fisher information is impractical to calculate (e.g., survival analysis)
- If survival times $T_i \stackrel{\text{iid}}{\sim} \text{Exp}(\theta)$ are subject to right censoring, then the observed information is d/θ^2 while the expected information is $\mathbb{E}d/\theta^2$, where d is the number of uncensored events
- First of all, the expected number of uncensored events is usually completely unknown and depends on many things that are not of scientific interest
- Second, does it even matter? Suppose we got lucky and observed more events than expected . . . is it really relevant that we could have obtained a sample with much less information?