Likelihood-based inference

Patrick Breheny

September 29

Introduction

- In previous lectures, we constructed and plotted likelihoods and used them informally to comment on likely values of parameters
- Our goal for today is to make this more rigorous, in terms of quantifying coverage and type I error rates for various likelihood-based approaches to inference
- With the exception of extremely simple cases such as the two-sample exponential model, exact derivation of these quantities is typically unattainable for survival models, and we must rely on asymptotic likelihood arguments

The score statistic

- Likelihoods are typically easier to work with on the log scale (where products become sums); furthermore, since it is only relative comparisons that matter with likelihoods, it is more meaningful to work with derivatives than the likelihood itself
- Thus, we often work with the derivative of the log-likelihood, which is known as the *score*, and often denoted U:

$$U_X(\theta) = \frac{d}{d\theta}\ell(\theta|X)$$

- Note that
  - U is a random variable, as it depends on X
  - U is a function of θ
  - For independent observations, the score of the entire sample is the sum of the scores for the individual observations:

$$U = \sum_i U_i$$

Mean

- We now consider some theoretical properties of the score
- It is worth noting that there are some regularity conditions that $f(x|\theta)$ must meet in order for these theorems to work; we'll discuss these in greater detail a little later
- **Theorem**: $\mathbb{E}(U) = 0$
- Note that maximum likelihood can therefore be viewed as a method of moments estimator with respect to the score statistic

Variance

- **Theorem**:

$$\mathbb{V}U = -\mathbb{E}\{U'\}$$

- The variance of U is given a special name in statistics: it is called the *Fisher information*, the *expected information*, or simply the *information*
- For notation, I will use $\mathcal{I}$ to represent the Fisher information, and $\mathcal{I}_i$ to represent the contribution to the Fisher information coming from the ith subject; note that under independence, $\mathcal{I} = \sum_i \mathcal{I}_i$
- Like the score, the Fisher information is a function of θ, although unlike the score, it is not random, as the random variable X has been integrated out

Some examples

- **Example #1**: For the normal mean model,

$$\mathcal{I}_i = \frac{1}{\sigma^2};$$

  this makes sense: as the data becomes noisier, less information is contained in each observation
- In the above example, the information is free of both X and μ (the parameter of interest); in general both can appear in the information, which gives rise to a few different ways of working with the information in practice
- **Example #2**: For the Poisson distribution,

$$U_i' = -X_i\lambda^{-2}$$

Observed information

- The Fisher information is therefore

$$\mathcal{I}_i(\lambda) = \lambda^{-1}$$

- Here, taking the expectation was straightforward; in general, it can be complicated, and for survival data analysis in particular, typically involves the censoring mechanism
- A simpler alternative is to use the observed values of $\{X_i\}$ rather than their expectation; this is known as the *observed information* and will be denoted I
- In the Poisson example,

$$I(\lambda) = \lambda^{-2} \sum_i x_i$$

Asymptotic distribution

We have a sum of independent terms for which we know the mean and variance; we can therefore apply the central limit theorem:

$$\sqrt{n}\{\bar{U} - \mathrm{E}(U)\} \xrightarrow{\mathsf{d}} N(0, \mathcal{I}_i),$$

or equivalently,

$$\frac{1}{\sqrt{n}} U \xrightarrow{\mathsf{d}} N(0, \mathcal{I}_i),$$

Consistency and information

- **Proposition:** Any consistent estimator of the information can be used in place of $\mathcal{I}_i$ from the previous slide, and the result still holds
- Thus, all of the following results hold:

$$\mathcal{I}(\theta_0)^{-1/2}U \xrightarrow{\mathsf{d}} \mathrm{N}(0,1)$$
$$\mathcal{I}(\hat{\theta})^{-1/2}U \xrightarrow{\mathsf{d}} \mathrm{N}(0,1)$$
$$I(\theta_0)^{-1/2}U \xrightarrow{\mathsf{d}} \mathrm{N}(0,1)$$
$$I(\hat{\theta})^{-1/2}U \xrightarrow{\mathsf{d}} \mathrm{N}(0,1)$$

provided that $\hat{\theta}$ is a consistent estimator of the true value θ_0

Multiple parameters

- All of these results can be extended to the case where multiple parameters are involved; this will be essential for studying any sort of regression model
- The score is now defined as

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

  where $\nabla \ell(\boldsymbol{\theta}|\mathbf{x})$ is the *gradient* of the log-likelihood, and has elements $\frac{\partial}{\partial \theta_1}\ell(\boldsymbol{\theta}|\mathbf{x}), \frac{\partial}{\partial \theta_2}\ell(\boldsymbol{\theta}|\mathbf{x}), \ldots$
- Note that
  - The score is now a $p \times 1$ vector; to denote this I will often write the score vector as $\mathbf{u}$
  - Finding the MLE now involves solving the system of equations $\mathbf{u}(\boldsymbol{\theta}) = \mathbf{0}$

Multivariate extensions

- The score still has mean zero: $\mathbb{E}(\mathbf{u}) = \mathbf{0}$
- The variance of the score is still the information, $\mathbb{V}(\mathbf{u}) = \boldsymbol{\mathcal{I}}$, although the information $\boldsymbol{\mathcal{I}}$ is now a $p \times p$ covariance matrix
- It is still true that under independence $\mathbf{u} = \sum_i \mathbf{u}_i$ and $\boldsymbol{\mathcal{I}} = \sum_i \boldsymbol{\mathcal{I}}_i$
- We again have that $\boldsymbol{\mathcal{I}} = -\mathbb{E}(\nabla \mathbf{u})$, where $\nabla \mathbf{u}$ is a $p \times p$ matrix of second derivatives with i, jth element $\frac{\partial}{\partial \theta_i} \frac{\partial}{\partial \theta_j} \ell(\boldsymbol{\theta} | \mathbf{x})$; this matrix is referred to as the *Hessian* matrix

Multivariate CLT results

- Finally, it is still true that

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

  where $\mathbf{1}$ denotes the $p \times p$ identity matrix
- As before, any of $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0)$, $\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}})$, $\mathbf{I}(\boldsymbol{\theta}_0)$, or $\mathbf{I}(\hat{\boldsymbol{\theta}})$ can be used as the information and the result still holds
- From the above, we also have

$$\mathbf{u}^T\boldsymbol{\mathcal{I}}^{-1}\mathbf{u} \xrightarrow{\mathsf{d}} \chi^2_p$$

Remarks on the non-IID case

- In general, all of these extensions are straightforward to show; however, it is worth noting that applying the central limit is somewhat more complex in the non-IID case
- In particular, it is not enough that the score have finite mean and variance in order to apply the CLT; we must also have $\boldsymbol{\mathcal{I}}/n \to \bar{\boldsymbol{\mathcal{I}}} \neq \mathbf{0}_{p\times p}$ and

  $$\boldsymbol{\mathcal{I}}_i \boldsymbol{\mathcal{I}}^{-1} \to \mathbf{0}_{p\times p}$$

  for all i
- Essentially, this means that, since each observation no longer contributes the same information, we have an added requirement that no single observation can dominate the information

Inference: Introduction

- How can we use these results to carry out likelihood-based inference?
- It turns out that there are three widely used techniques for doing so: the *score*, *Wald*, and *likelihood ratio* methods
- For the remainder of this lecture, we will motivate these three approaches and then apply them to exponentially distributed survival data as an illustration of how they work

Score test

- The score test follows most directly from our earlier derivations
- Here, to test $H_0 : \theta = \theta_0$, we simply calculate

$$\frac{U(\theta_0)}{\sqrt{I(\theta_0)}}$$

  and then compare it to a standard normal distribution
- As always, by inverting this test at $\alpha = 0.05$, we can obtain 95% confidence intervals for θ
- Note that the score test, unlike the next two approaches we will consider, does not even require estimating θ

Wald approximation

- The score test was first proposed by C. R. Rao; an alternative approach, first proposed by Abraham Wald, relies on a Taylor series approximation to the score function about the MLE
- **Proposition:**

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

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

Wald result

- Thus,

$$\boldsymbol{\mathcal{I}}^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \dot{\sim} \mathrm{N}(\mathbf{0}, \mathbf{1}), \text{ or}$$
$$\hat{\boldsymbol{\theta}} \dot{\sim} \mathrm{N}(\boldsymbol{\theta}_0, \boldsymbol{\mathcal{I}}^{-1})$$

- The MLE is therefore
  - Approximately normal...
  - ...with mean equal to the true value of the parameter...
  - ...and variance equal to the inverse of the information
- Based on this result, we can easily construct tests and confidence intervals for $\boldsymbol{\theta}$
- For simplicity, the above result is stated in terms of $\boldsymbol{\mathcal{I}}$; in practice it is typical to use $\mathbf{I}(\hat{\boldsymbol{\theta}})$ in the Wald approach

LRT approximation

- Finally, we could also consider the asymptotic distribution of the likelihood ratio, originally derived by Samuel Wilks
- This approach also involves a Taylor series expansion, but here we approximate the log-likelihood itself about the MLE, as opposed to the score
- **Proposition:**

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

LRT result

- Thus,

$$2\{\ell(\hat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}_0)\} \dot{\sim} \chi^2_p$$

- Note that

$$\exp\{-\chi^2_{1,(1-\alpha)}/2\} = 0.15;$$

  this was the basis for choosing 15% as a cutoff for $L(\theta)/L(\hat{\theta})$ in our likelihood intervals

- It is worth pointing out, however, that a 15% cutoff for $L(\theta)/L(\hat{\theta})$ is only appropriate for the single parameter case; in general, the cutoff would need to change in order to account for the additional degrees of freedom in the problem

Regularity conditions

The score, Wald, and LRT approaches derived here are all asymptotically equivalent to each other, and all hold provided that certain regularity conditions are met:

- θ is not a boundary parameter (otherwise we can't take an approximation about it)
- The information matrix $\boldsymbol{\mathcal{I}}_i(\theta_0)$ is finite and positive definite
- We can take up to third derivatives of $\int f(x|\theta)$ inside the integral, at least in the neighborhood of θ_0
- The distributions $\{f(x|\theta)\}$ have common support and are identifiable

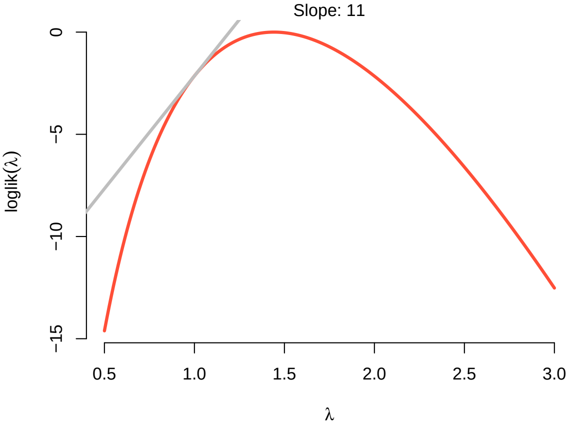
Reparameterization

- It is worth noting that the score and Wald approaches will be affected by reparameterization
- For example, if we decide to carry out inference for the log-hazard $\gamma = \log(\lambda)$ of an exponentially distributed time-to-event, we will obtain different score and Wald confidence intervals than if we constructed intervals for λ and then transformed them
- The likelihood ratio approach, however, since it doesn't involve any derivatives, will be unaffected by such transformations

Pike rat example

- To illustrate these approaches and the geometry behind them, we'll apply them to the Pike rat data
- For the purposes of this illustration, we'll assume the data follow an exponential distribution (which is not actually a very good assumption here) under independent censoring
- Also, we'll just look at overall survival without respect to pretreatment regimen

Score approach: $H_0 : \lambda = 1$



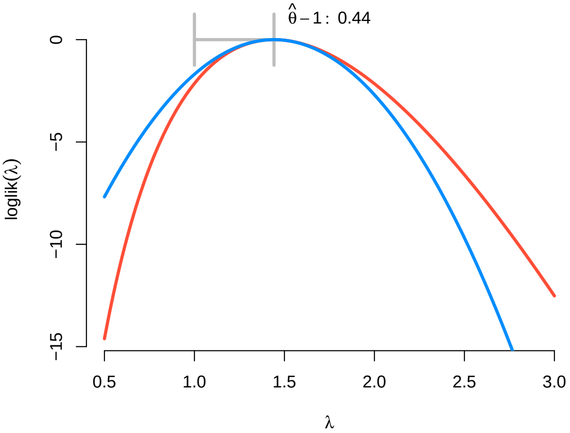
Score approach: $H_0 : \lambda = 1$ (cont'd)

- So, we observe a score of $d - v = 11$
- We would expect the score to be zero (i.e, if $\lambda = 1$, we'd expect to be near the top of the curve, where it's flat)
- Still, the standard error of the slope is $\sqrt{d} = 6$, so our observed score is only

$$Z = 11/6 = 1.84$$

standard deviations away from the mean, implying that we have insufficient evidence to rule out $\lambda = 1\,(p = 0.07)$

Wald approach: $H_0 : \lambda = 1$



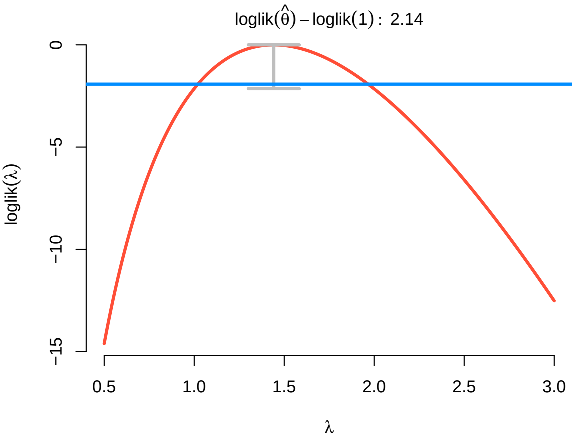
Wald approach: $H_0 : \lambda = 1$ (cont'd)

- So, we observe a difference of $\hat{\lambda} - \lambda_0 = d/v - 1 = 0.44$
- We would expect this difference to be near zero if λ was truly equal to 1
- However, the standard error $\hat{\theta}$ is $\sqrt{d}/v = 0.24$, so our observed difference is only

$$Z = 0.44/0.24 = 1.84;$$

  in this particular case, the score and Wald approaches coincide, but this is not true in general

Likelihood ratio approach: $H_0 : \lambda = 1$



Likelihood ratio approach: $H_0 : \lambda = 1$ (cont'd)

- So, we observe a difference of $\ell(\hat{\lambda}) - \ell(\lambda_0) = 2.14$
- Our p-value is therefore the area to the right of $2(2.14) = 4.29$ for a χ^2_1 distribution
- This turns out to be $p = 0.04$; thus, $\lambda = 1$ would be excluded from our likelihood ratio confidence interval despite being included in both the score and Wald intervals

“Exact” result

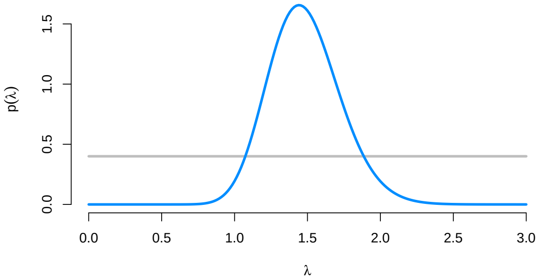
- For the exponential distribution, we could carry out something of an “exact” test based on the gamma distribution
- Here, our (one-sided) p-value would be the area to the left of V for a gamma distribution with shape parameter d and rate parameter λ_0, although it would only be exact in the case of type II censoring
- Nevertheless, the resulting one-sided p-value is 0.02; this is in good agreement with the two-sided p-value of 0.04 we got from the likelihood ratio test

Accuracy

- This small anecdote doesn't necessarily prove anything; nevertheless, it is the case the the likelihood ratio approach is typically the most accurate of the three
- To see why, consider analyzing a transformation, $g(\theta)$
- Some transformations will make the normal approximations for the score and Wald approaches more accurate (and some will make them less accurate)
- Suppose there exists a "best" transformation g^*; you could improve your score/Wald accuracy by finding and then applying g^*, but with the likelihood ratio test, you've already achieved that accuracy without even finding g^*

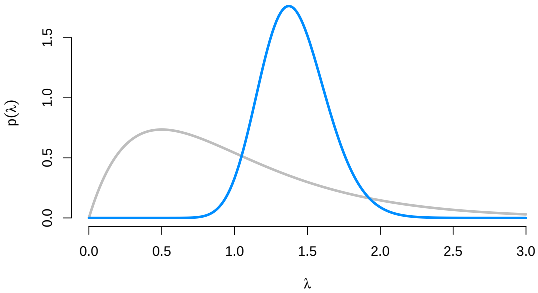
Bayesian approach: Uniform prior

We might also compare these results to the Bayesian approach, which doesn't require asymptotic approximations but does require the specification of a prior:



$$\mathbb{P}(\lambda < 1|d, v) = 0.014$$

Bayesian approach: Gamma(2,2) prior



$$\mathbb{P}(\lambda < 1 | d, v) = 0.026$$