Likelihood-based inference: Multiple parameters

Patrick Breheny

September 28

Multiple parameters

- All of the results we derived last time 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
- For the results that follow, we have the added regularity condition in the multivariate case that $\boldsymbol{\mathcal{I}}$ is not singular (i.e., that $\boldsymbol{\mathcal{I}}^{-1}$ exists)

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

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

  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

Multivariate CLT results

- Assuming this is satisfied, it is still true that

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

  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

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

Score, Wald, and LR tests

As in the univariate case, we can use this CLT and various Taylor series expansions to derive various tests of $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ by calculating

- Score:

$$\mathbf{u}(\boldsymbol{\theta}_0)^T \mathbf{I}(\boldsymbol{\theta}_0)^{-1} \mathbf{u}(\boldsymbol{\theta}_0)$$

- Wald:

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \mathbf{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$$

- Likelihood ratio:

$$2\{\ell(\hat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}_0)\}$$

and comparing the test statistic to a χ^2_p distribution

Nuisance parameters

- In practice, however, testing multivariate hypotheses like this is rare
- Instead, we typically wish to carry out inference regarding a single parameter of interest, θ_j, regardless of what the other parameters happen to be
- In this context, the other parameters $\boldsymbol{\theta}_{-j}$ are referred to as *nuisance parameters*; they are not the focus of the inference, but they must be properly accounted for in order to carry out inference on the quantity we are interested in

Wald approach

- Let's begin by seeing how nuisance parameters affect the Wald test
- The Wald approach is based on the result

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \dot{\sim} \mathrm{N}(\mathbf{0}, \mathbf{I}(\hat{\boldsymbol{\theta}})^{-1}),$$

and thus, marginally, we have

$$\hat{\theta}_j - \theta_j^* \dot{\sim} \mathrm{N}(0, [\mathbf{I}(\hat{\boldsymbol{\theta}})^{-1}]_{jj}),$$

or

$$\frac{\hat{\theta}_j - \theta_j^*}{\sqrt{[\mathbf{I}(\hat{\boldsymbol{\theta}})^{-1}]_{jj}}} \dot{\sim} \mathrm{N}(0, 1)$$

Impact of nuisance parameters

- Note, however, that

$$\frac{\hat{\theta}_j - \theta_j^*}{\sqrt{[\mathbf{I}(\hat{\boldsymbol{\theta}})^{-1}]_{jj}}} \neq \frac{\hat{\theta}_j - \theta_j^*}{\sqrt{[\mathbf{I}(\hat{\boldsymbol{\theta}})_{jj}]^{-1}}}$$

- In other words, the result we obtain from the Wald approach is *not* the same as simply ignoring the other parameters
- In particular, $[\mathbf{I}(\hat{\boldsymbol{\theta}})^{-1}]_{jj} \geq [\mathbf{I}(\hat{\boldsymbol{\theta}})_{jj}]^{-1}$; i.e., the standard error is always larger after accounting for nuisance parameters (or possibly stays the same)

Inverse of a partitioned matrix

- To see this, consider a matrix $\mathbf{M}$ partitioned as follows:

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \\ \mathbf{B}^T & \mathbf{D} \end{bmatrix}$$

- Then

$$\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{F}\mathbf{E}^{-1}\mathbf{F}^T & -\mathbf{F}\mathbf{E}^{-1} \\ \\ -\mathbf{E}^{-1}\mathbf{F}^T & \mathbf{E}^{-1} \end{bmatrix},$$

where $\mathbf{E} = \mathbf{D} - \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}$ and $\mathbf{F} = \mathbf{A}^{-1}\mathbf{B}$

Likelihood ratio and score approaches

- The effect of nuisance parameters on the score and LR tests is a bit more complicated
- Consider the problem of obtaining a likelihood ratio confidence interval for θ_j
- If θ_j was the only parameter, this is simply a root-finding problem in which we determine the values θ_j^L and θ_j^U where $2\{\ell(\hat{\theta}_j) - \ell(\theta_j)\} = \chi^2_{1,.95}$

The profile likelihood

- However, θ_j is not the only parameter, and in particular, if θ_j was restricted to equal θ_j^L, all the other MLEs would change as a consequence
- In other words, evaluating $\ell(\theta_j)$ is not simple, because it involves re-solving for $\hat{\boldsymbol{\theta}}_{-j}$ at every value of θ_j that we try out in our root-finding procedure
- The likelihood

$$L\{\theta_j, \hat{\boldsymbol{\theta}}_{-j}(\theta_j)\}$$

  is known as the *profile likelihood*, and the re-solving procedure is sometimes referred to as *profiling*
- Obtaining confidence intervals using either the score or likelihood ratio approaches involves profiling, but the Wald approach does not

Availability of LRCIs

- In practice, then, the tradeoff is that it is much faster and more convenient to obtain Wald CIs, since score and LR CIs involve profiling; however, likelihood ratio CIs tend to be more accurate
- Certainly, it is possible to write code that carries out profiling, and some software packages have implemented functions to do this for you (e.g., `glm`), but it is not as common as one would wish

Nuisance parameters in the Bayesian setting

- Finally, let's consider the Bayesian approach, which deals with nuisance parameters in a very different way
- In the Bayesian approach, inference is based on the posterior:

$$f(\boldsymbol{\theta}|\mathbf{x}) \propto f(\boldsymbol{\theta})f(\mathbf{x}|\boldsymbol{\theta})$$

- To obtain the marginal posterior for θ_j, then, we simply integrate over the possible values of $\boldsymbol{\theta}_{-j}$:

$$f(\theta_j|\mathbf{x}) = \int f(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta}_{-j}$$

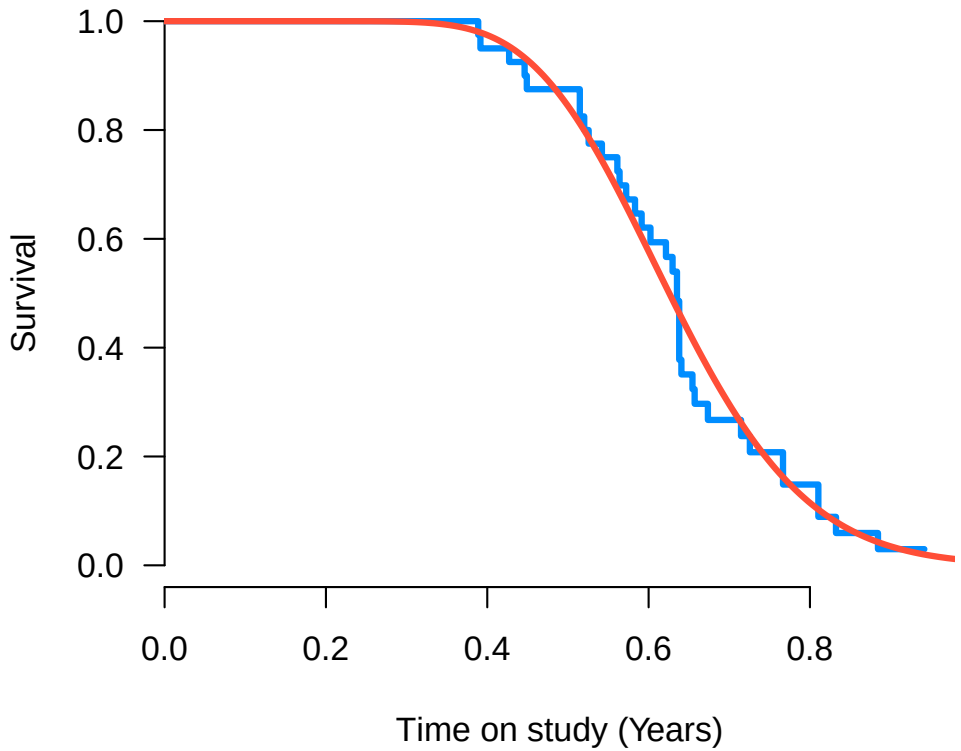
- In practice, we typically rely on numerical integration, generating a random sample from the full posterior and then just looking at the marginal distributions of interest

Pike rat data redux

- As a practical example, let's return to the Pike rat data from the previous lecture, but this time fit a Gamma distribution with shape parameter α and rate parameter λ to the data
- Since closed-form analytic solutions aren't available for this example, we will instead rely on numerical optimization, derivatives and integrals
- For our purposes, we'll consider λ to be our parameter of interest, and α a nuisance parameter

MLE

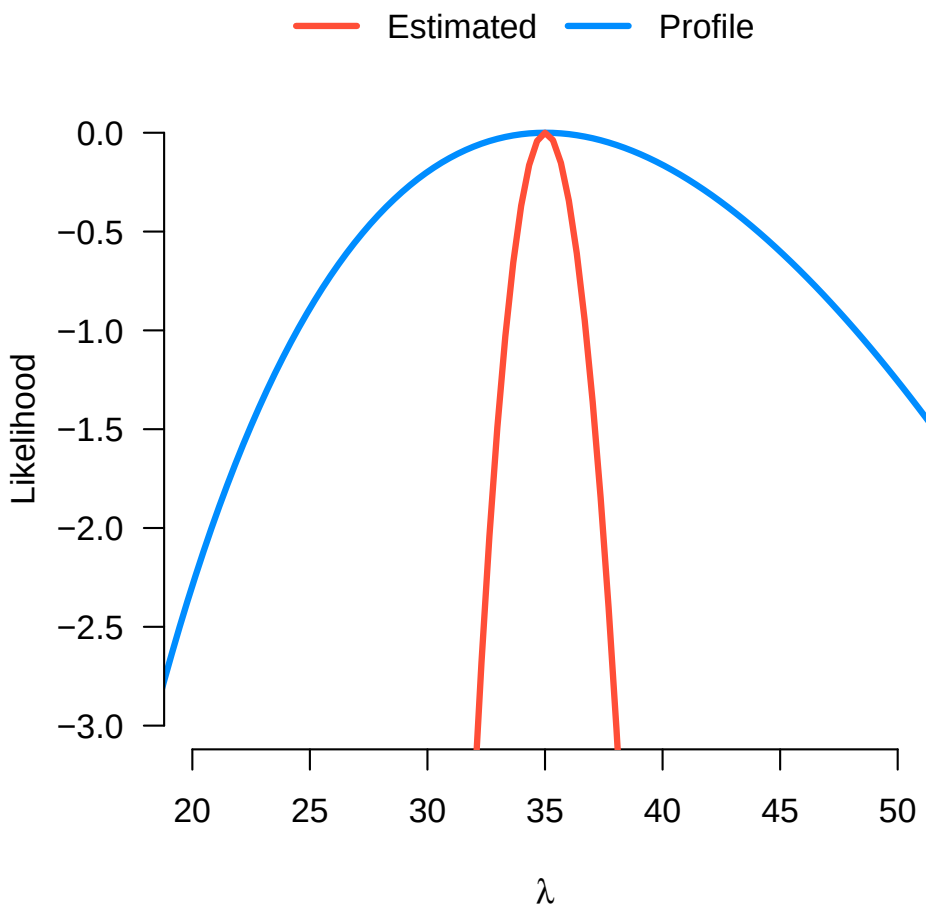
The MLE occurs at $\hat{\lambda} = 35$, $\hat{\alpha} = 22.2$, which provides a much better fit to the data than what we saw last time with the exponential distribution:



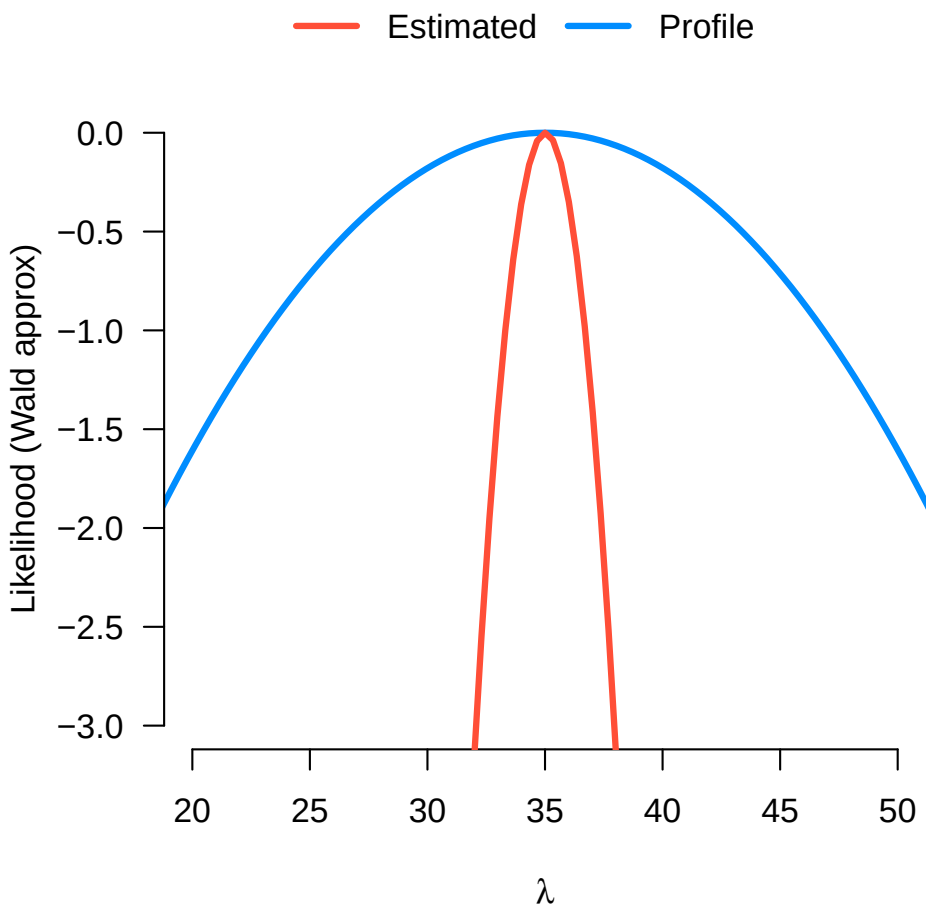
Profile vs. estimated likelihood

- To illustrate the impact of nuisance parameters upon inference in multiparameter problems, we will compare the profile likelihood to a simple, somewhat naïve likelihood called the "estimated likelihood"
- To construct the estimated likelihood, we will simple plug $\hat{\alpha} = 22.2$ into the likelihood and treat the likelihood as a single-parameter problem with respect to λ
- As we will see, this approach ignores uncertainty about α and results in unrealistic conclusions

Likelihoods



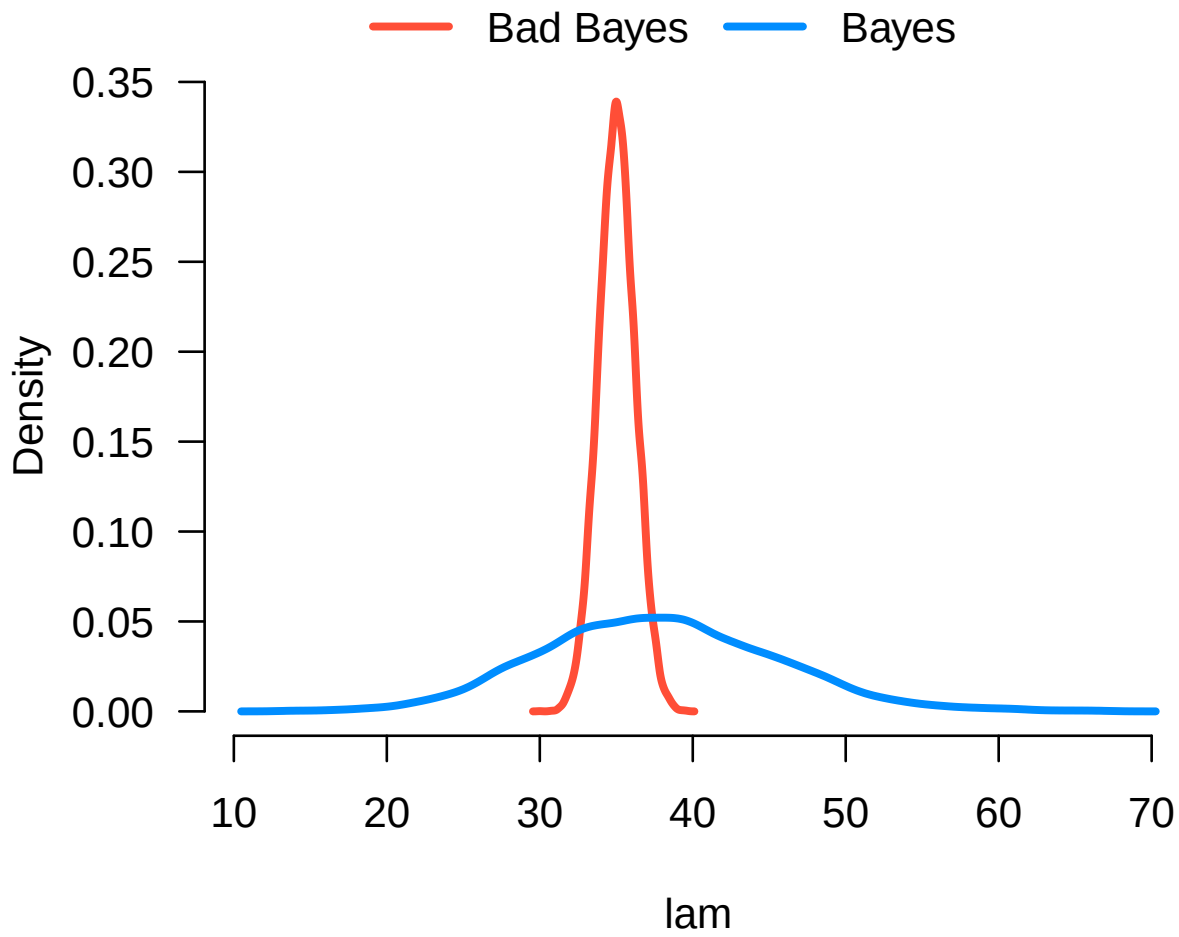
Wald approximation to likelihood



Bayesian inference

- A similar phenomenon happens in Bayesian inference
- Suppose that we simply replace α in the model with $\hat{\alpha}$ and treat $\hat{\alpha}$ as a constant (or, depending on your perspective, put a point prior with infinite strength on $\alpha = \hat{\alpha}$)
- This is known as an *empirical Bayes* approach
- Empirical Bayes certainly has its applications and can be a very useful statistical method, although this is an example of using it badly

Bayesian posterior for λ in the Pike rat study



Confidence/posterior intervals

	Nuisance parameters	
	Ignored	Accounted for
SE	1.2	8.4
Wald	(32.7, 37.4)	(18.6, 51.4)
Likelihood ratio	(32.7, 37.4)	(21.1, 54.1)
Bayes	(32.7, 37.4)	(23.6, 53.6)