Some properties of the MLE that we will explore are:

1. Equivariance: If $\hat{\theta}_n$ is the MLE of θ, then $g(\hat{\theta}_n)$ is the MLE of $g(\theta)$.

2. Consistency: $\hat{\theta}_n \overset{P}{\to} \theta_*$, where θ_* is the true value of the parameter.

3. Asymptotic normality: $(\hat{\theta}_n - \theta_*)/se(\hat{\theta}_n) \overset{D}{\to} N(0,1)$.

4. Asymptotic efficiency: The MLE has the smallest asymptotic variance among asymptotically normal estimators.

Conditions for the last three can be somewhat technical, so we'll start with the case that $\theta \in \Theta \subseteq \mathbb{R}$ and will focus more on intuition than on details.

Equivariance: Let $\tau = g(\theta)$ be a function of θ. Let $\hat{\theta}_n$ be the MLE of θ. Then $\hat{\tau}_n = g(\hat{\theta}_n)$ is the MLE of τ.

Proof: Suppose that g is one-to-one. Then it possesses an inverse g^{-1}, and we can define the induced likelihood $\mathcal{L}^*(\tau) = \mathcal{L}(g^{-1}(\tau))$. But for any τ,

$$\mathcal{L}^*(\tau) = \mathcal{L}(g^{-1}(\tau)) \leq \mathcal{L}(\hat{\theta}) = \mathcal{L}^*(g(\hat{\theta}))$$

so $\hat{\tau} = g(\hat{\theta})$ maximizes $\mathcal{L}^*$.

The general case is only slightly more complicated; we define

$$\mathcal{L}^*(\tau) = \sup_{\theta : g(\theta) = \tau} \mathcal{L}(\theta)$$

The following conditions are sufficient for consistency of the MLE:

1. $X_1, \ldots, X_n$ are iid with density $f(x;\theta)$.

2. Identifiability, i.e. if $\theta \neq \theta'$, then $f(x;\theta) \neq f(x;\theta')$.

3. The densities $f(x;\theta)$ have common support, i.e. $\{x : f(x;\theta) > 0\}$ is the same for all θ.

4. The parameter space Θ contains an open set ω of which the true parameter value θ_* is an interior point.

5. The function $f(x;\theta)$ is differentiable with respect to θ in ω.

These conditions ensure uniform convergence in probability of a normalized form of the log-likelihood to its expected value.

Note that

$$\begin{aligned}
\ell_n(\theta) &= \sum_{i=1}^{n} \log f(X_i;\theta) \\
&\propto \frac{1}{n}\sum_{i=1}^{n} \log f(X_i;\theta) \\
&\overset{P}{\to} E_{\theta_*}[\log f(X_1;\theta)] \text{ for any fixed } \theta \text{ by WLLN}
\end{aligned}$$

where θ_* denotes the true value of θ. Showing consistency requires that the convergence is uniform in θ. We also need to show that $E_{\theta_*}[\log f(X_1;\theta)]$ is maximized at $\theta = \theta_*$.

One class of distributions that satisfies the conditions is known as the exponential family. For $\Theta \subseteq \mathbb{R}$, these have densities that can be written as

$$f(x;\theta) = h(x)c(\theta)\exp\{\eta(\theta)T(x)\}$$

Example: Show that each belongs to the exponential family

- $Binomial(n,p)$ with n known
- $Exponential(\lambda)$

Show that $Unif(0,\theta)$ does not.

Define the score function $s(X;\theta) = \frac{\partial}{\partial\theta} \log f(X;\theta)$.

Then the Fisher information (based on n observations) is

$$\begin{aligned}
I_n(\theta) &= V_\theta\left(\frac{\partial}{\partial\theta}\ell_n(\theta)\right) \\
&= V_\theta\left(\sum_{i=1}^{n} s(X_i;\theta)\right) \\
&= \sum_{i=1}^{n} V_\theta(s(X_i;\theta)) \quad \text{(if } X_1,\ldots,X_n \text{ are independent)} \\
&= nV_\theta(s(X_1;\theta)) \quad \text{(if } X_1,\ldots,X_n \text{ are identically distributed)} \\
&= nI_1(\theta) \\
&\equiv nI(\theta)
\end{aligned}$$

In addition, under a condition satisfied for exponential family models, we can calculate

$$I(\theta) = -E_\theta \left[\frac{\partial^2}{\partial \theta^2} \log f(X;\theta) \right]$$

Example: Suppose $X_1, \ldots, X_n \overset{iid}{\sim} Pois(\lambda)$. Calculate $I_n(\lambda)$.

The “observed” Fisher information

$$I_n^{obs}(\theta) = \frac{-\partial^2}{\partial \theta^2} \sum_{i=1}^{n} \log f(X_i;\theta)$$

measures the curvature of the log-likelihood function. In particular $I_n^{obs}(\hat{\theta})$ measures the curvature at the MLE. The more peaked $\ell_n(\theta)$ is around $\hat{\theta}$, the more "information" the likelihood gives us. $I(\theta)$ measures the average value of this quantity.

Under two additional conditions (also satisfied by iid observations under exponential family models), we have

- Asymptotic normality: $\sqrt{n}(\hat{\theta}_n - \theta) \stackrel{D}{\to} N(0, 1/I(\theta))$

- Asymptotic efficiency: If $\tilde{\theta}_n$ is some other estimator s.t. $\sqrt{n}(\tilde{\theta}_n - \theta) \stackrel{D}{\to} N(0, v(\theta))$, then $v(\theta) \geq 1/I(\theta)$ for all θ.

Asymptotic normality still holds replacing $I(\theta)$ by $I(\hat{\theta})$, that is,

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sqrt{1/I(\hat{\theta}_n)}} \stackrel{D}{\to} N(0, 1)$$

We can use this to construct approximate $1 - \alpha$ confidence intervals for θ.

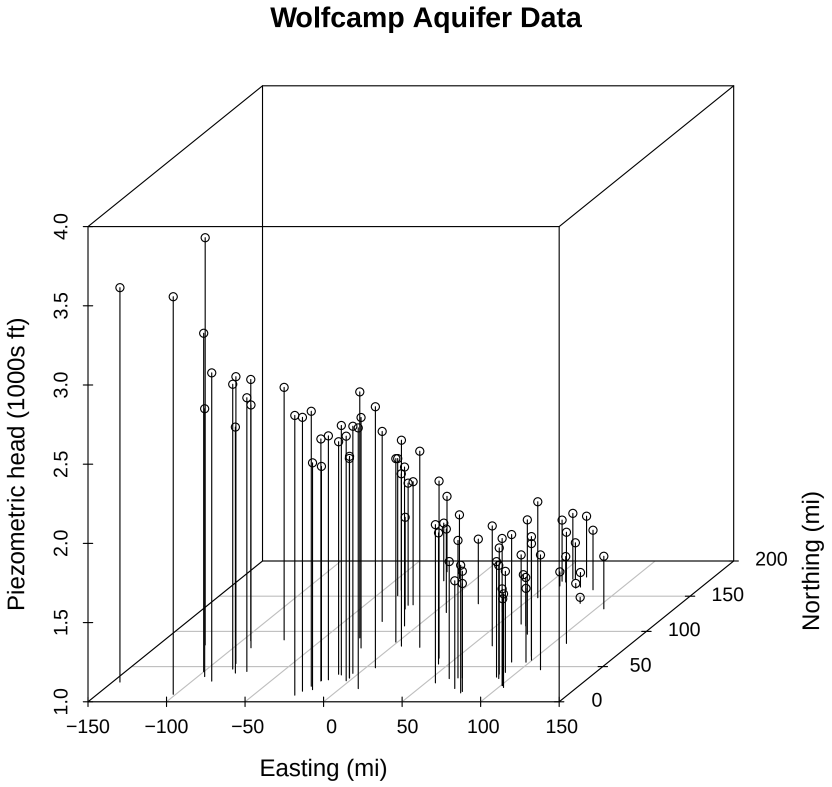
Under each of the following models, find the MLE for θ and calculate an approximate 95% confidence interval using the limiting normal distribution.

1. $X_1, \ldots, X_n \overset{iid}{\sim} Exp(\theta)$

2. $X_1, \ldots, X_n \overset{iid}{\sim} Binomial(m, \theta)$ for known m

3. $X_1, \ldots, X_n \overset{iid}{\sim} Normal(\theta, \sigma^2)$ for known σ^2

A motivating example for more complicated likelihood problems: Given measurements of hydraulic head from an aquifer, create a predicted surface.



The Wolfcamp Aquifer lies below Deaf Smith County, Texas, once under consideration by DOE as a nuclear waste repository site.

Creating a smooth surface from the measurements would allow us to predict the path of potential contaminants.

Aside: Multivariate normal distribution

The multivariate normal distribution for a vector $Z = (Z_1, Z_2, \ldots, Z_n)'$ with mean vector μ and covariance matrix Σ has pdf

$$f(z; \mu, \Sigma) = (2\pi)^{-1/2} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2}(z-\mu)'\Sigma^{-1}(z-\mu) \right\}$$

Let μ_i denote the i^{th} element of μ, and Σ_{ij} the element of Σ in the i^{th} row and j^{th} column. Then

- $Z_i \sim N(\mu_i, \Sigma_{ii})$
- $Cov(Z_i, Z_j) = \Sigma_{ij}$

In the aquifer example, we could fit a universal kriging model, which is just a multivariate normal model where μ and Σ have special structure.

In this example, we could take

$$\mu_i = E[Z_i] = \beta_0 + \beta_1 x_i + \beta_2 y_i$$

where (x_i, y_i) is the location of observation i.

The observations are clearly not independent, so Σ is not diagonal. One model would be to have correlation decay with distance, such as

$$\Sigma_{ij} = Cov(Z_i, Z_j) = \sigma^2 \exp\{-d_{ij}/\rho\}$$

where $d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$.

There is no closed form expression for the MLE of $\theta = (\beta_0, \beta_1, \beta_2, \sigma^2, \rho)$.

When $\theta = (\theta_1, \ldots, \theta_k)$, we define the Fisher information matrix as follows.

The Hessian matrix is the matrix of second partial derivatives of the log-likelihood, with

$$H_{jj} = \frac{\partial^2}{\partial \theta_j^2} \ell_n(\theta); \quad H_{jk} = \frac{\partial^2}{\partial \theta_j \partial \theta_k} \ell_n(\theta)$$

The Fisher information matrix is

$$I_n(\theta) = - \left[\begin{array}{ccc} E_\theta(H_{11}) & \cdots & E_\theta(H_{1k}) \\ E_\theta(H_{21}) & \cdots & E_\theta(H_{2k}) \\ \vdots & \vdots & \vdots \\ E_\theta(H_{k1}) & \cdots & E_\theta(H_{kk}) \end{array} \right]$$

Let $\hat{\theta}$ be the (vector valued) MLE, and let $J_n(\theta) = I_n(\theta)^{-1}$. Then under appropriate regularity conditions and for large n,

$$\hat{\theta}_n \overset{D}{\approx} N(\theta, J_n(\theta))$$

We can use the marginal densities ($\hat{\theta}_{n,i} \overset{D}{\approx} N(\theta_i, J_{n,ii}(\theta))$) to construct 95% confidence intervals for the individual parameters.

Example: Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. The MLEs for μ and σ are $\hat{\mu}_n = \bar{X}_n$ and $\hat{\sigma}_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$. In addition...

$$I_n(\mu,\sigma) = \left[\begin{array}{cc} n/\sigma^2 & 0 \\ 0 & 2n/\sigma^2 \end{array} \right]$$

$$J_n(\mu,\sigma) = I_n(\mu,\sigma)^{-1} = \left[\begin{array}{cc} \sigma^2/n & 0 \\ 0 & \sigma^2/(2n) \end{array} \right]$$

Using the fact that both $\hat{\mu}_n$ and $\hat{\sigma}_n$ are consistent, we can plug in to get

$$\hat{X}_n \pm 2\sqrt{\frac{\hat{\sigma}^2}{n}} \text{ and } \hat{\sigma}_n \pm 2\sqrt{\frac{\hat{\sigma}^2}{2n}}$$

as approximate 95% confidence intervals for μ and σ.

Multiparameter delta method

Suppose $\tau = g(\theta_1, \ldots, \theta_k)$ is a differentiable function. Let $\nabla g = (\frac{\partial}{\partial \theta_1} g(\theta) \cdots \frac{\partial}{\partial \theta_k} g(\theta))'$ be the gradient of g and suppose that ∇g evaluated at $\hat{\theta}_n$ is not zero. Then

$$\frac{\hat{\tau}_n - \tau}{\widehat{se}(\hat{\tau}_n)} \overset{D}{\to} N(0,1)$$

where

$$\widehat{se}(\hat{\tau}_n) = \sqrt{(\hat{\nabla} g)' J_n(\hat{\theta}_n)(\hat{\nabla} g)}$$

and $\hat{\nabla} g$ is ∇g evaluated at $\hat{\theta}_n$.

Example: Continuing the last example, let $\tau = g(\mu, \sigma) = \mu/\sigma$. Find the MLE for τ and its limiting normal distribution.

In many cases, it's not possible to find a closed-form expression for the MLE in multiparameter models. This is true even for some common distributions like the Gamma and Beta distributions.

However, numerical optimization is a highly developed field that comes to our rescue in applied problems (that is, when we have actual values for $X_1, \ldots, X_n$).

Most of these algorithms are written for minimization, so we need to

- Write a function for the negative log-likelihood
- Minimize it numerically
- Examine the behavior of the negative log-likelihood at the minimum
- Optionally, get a numerical approximation of the Hessian and compute the observed Fisher information matrix

In these problems, we often replace the Fisher information matrix with the observed Fisher information matrix. In many cases confidence intervals constructed using the observed Fisher information actually perform better than those using the Fisher information, so this is not a big issue.

See `betaexample.R` and `aquifer.R` for examples of this process in a toy dataset and for the aquifer example.