Parametric Inference

A parametric model has the form

$$\mathcal{F} = \{F(x;\theta) : \theta \in \Theta\}$$

where $\Theta \subseteq \mathbb{R}^k$ is the parameter space.

We typically choose a class $\mathcal{F}$ based on knowledge about the particular problem. We might say we're making certain assumptions about the data generating mechanism. It's good practice when using a parametric model to look for violations of these assumptions.

We'll begin with two methods for constructing estimators of θ: the method of moments and maximum likelihood estimation.

Suppose $\theta = (\theta_1, \ldots, \theta_k)$. For $j = 1, \ldots, k$, define the j^{th} moment

$$\alpha_j \equiv \alpha_j(\theta) = E_\theta[X^j] = \int x^j dF_\theta(x)$$

and the j^{th} sample moment $\hat{\alpha}_j = \frac{1}{n}\sum_{i=1}^n X_i^j$.

The method of moments estimator $\hat{\theta}_n$ is defined to be the value of θ s.t.

$$\begin{aligned} \alpha_1(\hat{\theta}_n) &= \hat{\alpha}_1 \\ \alpha_2(\hat{\theta}_n) &= \hat{\alpha}_2 \\ \vdots \quad &\vdots \quad \vdots \\ \alpha_k(\hat{\theta}_n) &= \hat{\alpha}_k \end{aligned}$$

MOM generalization: Instead of using $\alpha_j(\theta) = E_\theta[X^j]$, we can consider $\alpha_j(\theta) = E_\theta[g(X)^j]$ and find $\hat{\theta}_n$ s.t. $\alpha_j(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^n g(X_i)^j$, $j = 1, ..., k$.

The maximum likelihood estimator (MLE) is obtained by maximizing the likelihood function

$$\begin{aligned}\mathcal{L}_n(\theta) &= f(X_1, \ldots, X_n; \theta) \\ &= \prod_{i=1}^{n} f(X_i; \theta) \quad \text{if the data are independent}\end{aligned}$$

That is, the likelihood is just the joint density of the data, but viewed as a function of θ.

It's often easier to work with the log-likelihood function

$$\ell_n(\theta) = \log \mathcal{L}_n(\theta) = \sum_{i=1}^{n} \log f(X_i; \theta)$$

If the log-likelihood is differentiable with respect to θ, possible candidates for the MLE are those in the interior of Θ that solve

$$\frac{\partial}{\partial \theta_j} \ell_n(\theta) = 0, \quad j = 1, \ldots, k$$

We still need to check that we've found the global maximum. Also note that if the maximum occurs on the boundary of Θ, the first derivative may not be zero.

It's not always possible to maximize the likelihood analytically, and in these cases we turn to numerical maximization methods.

Examples:

- Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$. Find the MLE for θ.
- Now solve the same problem, but with the restriction $\Theta = [0, \infty)$.
- Let $X_1, \ldots, X_n \overset{iid}{\sim} Unif(0, \theta]$. Find the MLE and the MOM for θ.
- Let $X_1, \ldots, X_n \overset{iid}{\sim} Unif[\theta, \theta + 1]$. Find the MLE for θ.
- Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$. Show that MOM and MLE are equivalent for this distribution family. Can this result be generalized?

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: $\hat{\theta}_n \xrightarrow{P} \theta^*$, where θ^* is the true value of the parameter.

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

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,

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

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

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

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 $\Theta \subseteq \mathbb{R}^d$, these have densities that can be written as

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

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

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

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

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

ASIDE: Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with mean μ and variance $\sigma^2 > 0$. Define the sample mean

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Then, as $n \to \infty$, the standardized random variable

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}}$$

converges in distribution to a standard normal random variable.

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

Summary

For an exponential family,

1. MOM and MLE are equivalent (check by yourself).

2. MLE is consistent.

3. MLE is asymptotically normal after an appropriate linear transformation.

4. MLE is asymptotically efficient (smallest asymptotic variance defined through fisher information).

Fisher Information matrix

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

$$I_n(\theta) = - \begin{bmatrix} E_\theta(H_{11}) & \cdots & E_\theta(H_{1k}) \\ E_\theta(H_{21}) & \cdots & E_\theta(H_{2k}) \\ \vdots & \vdots & \vdots \\ E_\theta(H_{k1}) & \cdots & E_\theta(H_{kk}) \end{bmatrix}$$

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

$$\sqrt{n}(\hat{\theta}_n - \theta) \overset{D}{\approx} N(0, nJ_n(\theta))$$

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...

$$I_n(\mu, \sigma) = \begin{bmatrix} n/\sigma^2 & 0 \\ 0 & 2n/\sigma^2 \end{bmatrix}$$

$$J_n(\mu, \sigma) = I_n(\mu, \sigma)^{-1} = \begin{bmatrix} \sigma^2/n & 0 \\ 0 & \sigma^2/(2n) \end{bmatrix}$$

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

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

as approximate 95% confidence intervals for μ and σ.

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

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

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

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

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

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

where

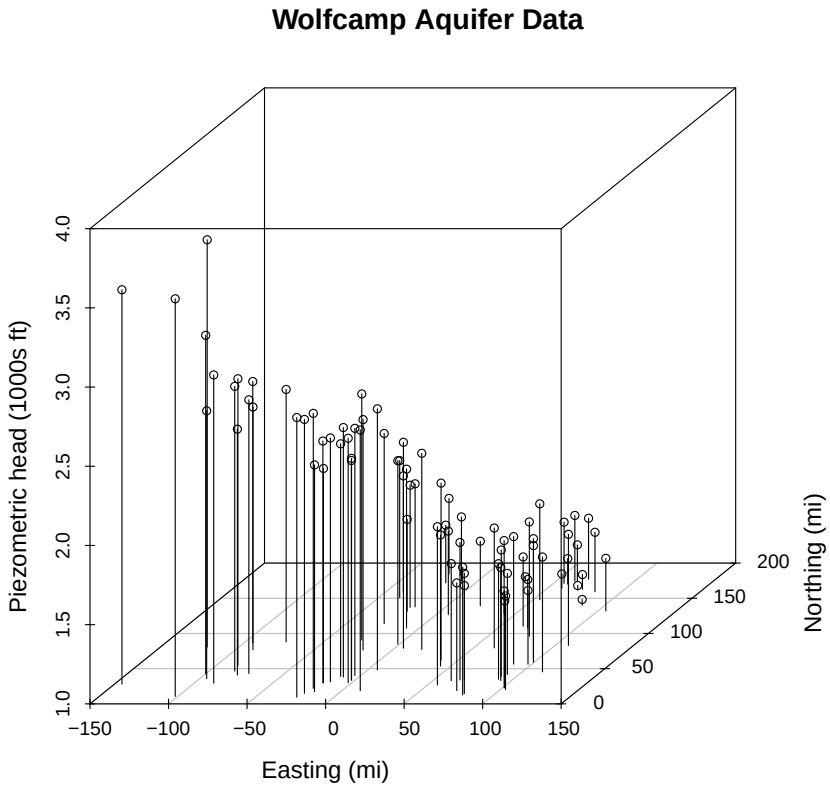
$$\widehat{se}(\hat{\tau}_n) = \sqrt{(\hat{\nabla} g)' J_n(\hat{\theta}_n)(\hat{\nabla} g)}$$

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

Example: Continuing the example on page 19, let $\tau = g(\mu, \sigma) = \mu/\sigma$. Find the MLE for τ and its limiting normal distribution.

A more complicated likelihood problem

Given measurements of hydraulic head from an aquifer, how to 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.

In the aquifer example, we could fit 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 $\theta = (\beta_0, \beta_1, \beta_2, \sigma^2, \rho)$.

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