STAT 528 - Advanced Regression Analysis II

Exponential family theory (part 2)

Daniel J. Eck
Department of Statistics
University of Illinois

Last time

Let Y be a regular full exponential family in canonical form. Then Y has log likelihood given by

$$l(\theta) = \langle Y, \theta \rangle - c(\theta),$$

and

$$\begin{aligned} \mathrm{E}_\theta(Y) &= \nabla c(\theta), \\ \mathrm{Var}_\theta(Y) &= \nabla^2 c(\theta). \end{aligned}$$

Examples include: Binomial, Poisson, Normal, etc.

Learning Objectives Today

Mean value parameters

The mean of the canonical statistic $\mathsf{E}_\theta(Y)$ is also a parameter.

It is given as a function g of the canonical parameter θ,

$$g(\theta) = \nabla c(\theta) = \mathsf{E}_\theta(Y) = \mu. \tag{1}$$

We will refer to $g : \theta \to \mu$ as the change-of-parameter map (or change-of-parameter) from canonical parameter θ to mean value parameter μ.

Example: Binomial

Recall that the log likelihood for the binomial distribution (after dropping terms)

$$y \log(p) + (n - y) \log(1 - p)$$

can be written in canonical form with

$$c(\theta) = n \log(1 + \exp(\theta)),$$

where $\theta = \log(p/(1 - p))$ and $p = \exp(\theta)/(1 = \exp(\theta))$. Now,

$$\nabla c(\theta) = n \left(\frac{\exp(\theta)}{1 + \exp(\theta)} \right) = np,$$

and this is $E_\theta(Y)$.

Mean value parameters

Theorem

For a regular full exponential family, the change-of-parameter from canonical to mean value parameter is invertible if the model is identifiable. Moreover both the change-of-parameter and its inverse are infinitely differentiable.

See notes for proof.

Mean value parameters

Recall from last time that an exponential family is identifiable if and only if the canonical statistic is NOT concentrated on a hyperplane

$$H = \{y : y^T v = a\} \tag{2}$$

Theorem

An exponential family fails to be identifiable if and only if the canonical statistic is concentrated on a hyperplane. If that hyperplane is given by (2) *and the family is full, then* θ *and* $\theta + sv$ *are in the full canonical parameter space and correspond to the same distribution for every canonical parameter value* θ *and every scalar* s*.*

Multivariate monotone

The change-of-parameter mapping $g : \theta \to \mu$ is multivariate monotone,

$$(g(\theta_1) + g(\theta_2))'(\theta_1 - \theta_2) \geq 0, \tag{3}$$

where $g(\theta_i) = \mu_i$, $i = 1, 2$.

If we rewrite (3) using subscripts and consider θ_1 and θ_2 that differ in only one coordinate, say the kth, then we get

$$(\mu_{1k} - \mu_{2k})(\theta_{1k} - \theta_{2k}) > 0,$$

which says *if we increase one component of the canonical parameter vector, leaving the other components fixed, then the corresponding component of the mean value parameter vector also increases, and the other components can go any which way*.

MLE

The derivative of the log likelihood is

$$\nabla l(\theta) = y - \nabla c(\theta).$$

The second derivative is

$$\nabla^2 l(\theta) = -\nabla^2 c(\theta).$$

Hence observed and expected Fisher information for the canonical parameter vector θ are the same

$$I(\theta) = \nabla^2 c(\theta). \tag{4}$$

Note: Fisher information measures the expected curvature of the log likelihood around the true parameter value.

If the likelihood is sharply curved around θ – the expected information $I(\theta)$ is large – then a small change in θ can lead to a drastic decrease in the likelihood. Conversely, if $I(\theta)$ is small then small changes in θ will not affect the likelihood that much.

These heuristics are important when we cover data separation.

When

- the model is identifiable
- the canonical statistic vector Y is not concentrated on a hyperplane
- the second derivative is negative definite everywhere.

Hence the log likelihood is strictly concave, hence the maximum likelihood estimate is unique if it exists. Thus,

$$y = \nabla c(\hat{\theta}),$$

and

$$\hat{\theta} = g^{-1}(y).$$

Observed equals expected

The MLE, if it exists, must be a point where the first derivative is zero, that is, a θ that satisfies

$$y = \nabla c(\theta) = \mathrm{E}_\theta(Y).$$

The MLE is the (unique if the model is identifiable) parameter value that makes the observed value of the canonical statistic equal to its expected value.

We call this the **observed equals expected** property of maximum likelihood in exponential families.

This property is even simpler to express in terms of the mean value parameter. By invariance of maximum likelihood under change-of-parameter, the MLE for μ is

$$\hat{\mu} = g(\hat{\theta}) = \nabla c(\hat{\theta})$$

and the observed equals expected property is therefore

$$y = \hat{\mu}. \tag{5}$$

Non-Existence of the MLE

We cannot prove the maximum likelihood estimate (for the canonical parameter) exists.

Consider the binomial distribution. The MLE for the usual parameterization is $\hat{p} = y/n$.

The canonical parameter is $\theta = \text{logit}(p)$.

But $\hat{\theta} = \text{logit}(\hat{p})$ does not exist when $\hat{p} = 0$ or $\hat{p} = 1$, which is when we observe zero successes or when we observe n successes in n trials.

We will revisit this topic when we discuss GLMs.

IID Data

Suppose $y_1, \ldots, y_n$ are independent and identically distributed (IID) from some distribution in an exponential family.

The log likelihood for sample size n is

$$l_n(\theta) = \sum_{i=1}^{n} [\langle y_i, \theta \rangle - c(\theta)] = \langle \sum_{i=1}^{n} y_i, \theta \rangle - nc(\theta), \tag{6}$$

and we see that the above log likelihood is an exponential family with:

- canonical statistic $\sum_{i=1}^{n} y_i$,
- cumulant function $\theta \mapsto nc(\theta)$, and
- canonical parameter θ and full canonical parameter space Θ the same as the originally given family from which every observation is a member.

Thus IID sampling gives us a new exponential family, but still an exponential family.

Asymptotics of MLE for regular full exponential family

Rewrite (6) as

$$l_n(\theta) = n\left[\langle \bar{y}_n, \theta \rangle - c(\theta)\right]$$

so that

$$\nabla l_n(\theta) = n\left[\bar{y}_n - \nabla c(\theta)\right].$$

From which we see that for an identifiable regular full exponential family where the MLE must be a point where the first derivative is zero, we can write

$$\begin{aligned} \nabla l_n(\theta) = n\left[\bar{y}_n - \nabla c(\theta)\right] &= 0, \\ \bar{y}_n - \nabla c(\hat{\theta}) &= 0. \end{aligned}$$

And this yields

$$\bar{y}_n = \nabla c(\hat{\theta}).$$

Recall the change-of-parameters mapping $g : \theta \mapsto \mu$ is invertible when the model is identifiable. We can write

$$\hat{\theta}_n = g^{-1}(\bar{y}_n), \tag{7}$$

when the MLE exists.

When the MLE does not exist, $\bar{y}_n$ is not in the domain of g^{-1}.

By the multivariate central limit theorem (CLT)

$$\sqrt{n}\,(\bar{y}_n - \mu) \to N\,(0, I(\theta))$$

and we know that g^{-1} is differentiable (Theorem 1 in this slide deck) with the derivative given by

$$\nabla g^{-1}(\mu) = [\nabla g(\theta)]^{-1},$$

where

$$\mu = g(\theta) \quad \text{and} \quad \theta = g^{-1}(\mu).$$

So the usual asymptotics of maximum likelihood

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \to N\left(0, I(\theta)^{-1}\right) \tag{8}$$

is just the multivariate delta method applied to the multivariate CLT.

Example: Bernoulli distribution

Let $y_1, \ldots, y_n \overset{iid}{\sim} \text{Bernoulli}(p)$ where $0 < p < 1$.

Then

$$l_n(\theta) = \langle \sum_{i=1}^{n} y_i, \theta \rangle - nc(\theta)$$

where

$$g(\theta) = \frac{\exp(\theta)}{1 + \exp(\theta)} = p$$

$$g^{-1}(p) = \log\left(\frac{1}{1-p}\right) = \theta,$$

and

$$\hat{p} = \bar{y}_n.$$

We have

$$\sqrt{n}\,(\bar{y}_n - p) \to N\,(0, I(\theta))\,,$$

where

$$\begin{aligned} I(\theta) &= \nabla^2 c(\theta) \\ &= \nabla^2 \log(1 + \exp(\theta)) \\ &= \nabla \left(\frac{\exp(\theta)}{1 + \exp(\theta)} \right) \\ &= \left(\frac{\exp(\theta)}{1 + \exp(\theta)} \right) \left(\frac{1}{1 + \exp(\theta)} \right) \\ &= p(1 - p). \end{aligned}$$

We have,

$$\begin{aligned}\sqrt{n}\left(\hat{\theta}_n - \theta\right) &= \sqrt{n}\left(g^{-1}(\bar{y}_n) - g^{-1}(p)\right) \\ &\to N\left(0, \nabla g^{-1}(p) I(\theta) \left[\nabla g^{-1}(p)\right]^T\right) \\ &= N\left(0, \left[\nabla g(\theta)\right]^{-1} I(\theta) \nabla \left[\nabla g(\theta)\right]^{-1}\right),\end{aligned}$$

where

$$\left[\nabla g(\theta)\right]^{-1} = \left[\nabla^2 c(\theta)\right]^{-1} = I(\theta)^{-1}.$$

Therefore,

$$\begin{aligned}\sqrt{n}\left(\hat{\theta}_n - \theta\right) &\to N\left(0, I(\theta)^{-1}\right) \\ &= N\left(0, \frac{1}{p(1-p)}\right).\end{aligned}$$