STAT 528 - Advanced Regression Analysis II

Exponential family theory (part 4)

Daniel J. Eck
Department of Statistics
University of Illinois

Last time

- generalized linear models (GLMs)
- different parameterizations
- motivation of logistic regression
- inference for model parameters
- comparing models

Learning Objectives Today

- Optimization
- Sufficiency
- Maximum Entropy
- Wrap-up

Background

We start with a regular full exponential family with log likelihood

$$\langle y, \theta \rangle - c(\theta)$$

where $\theta, y \in \mathbb{R}^n$. We have a canonical linear submodel where

- $\theta = M\beta$
- $\beta \in \mathbb{R}^p$
- $p < n$

and log likelihood

$$\langle M'y, \beta \rangle - c_\beta(\beta).$$

A depiction of the transformations necessary to change between parameterizations.

$$\begin{tikzcd}
\theta \arrow[r, shift left] \arrow[d, shift left, "\mu = \nabla c(\theta)"] & \beta \arrow[l, shift left, "\theta = M\beta"] \arrow[d, shift left, "\tau = \nabla_{\beta} c(M\beta)"] \\
\mu \arrow[u, shift left] \arrow[r, shift left, "\tau = M^T \mu"] & \tau \arrow[l, shift left] \arrow[u, shift left]
\end{tikzcd}$$

The only way to determine the MLEs is to maximize the log likelihood

$$\langle M'y, \beta \rangle - c_\beta(\beta).$$

to obtain $\hat{\beta}$ and then

- $\hat{\theta} = M\hat{\beta}$
- $\hat{\mu} = \nabla c(\hat{\theta})$
- $\hat{\tau} = M^T \hat{\mu}$.

Our goal is to compute

$$\operatorname{argmax}_{\beta} \left(\langle M'y, \beta\rangle - c_{\beta}(\beta)\right)$$

We will discuss a few ways of computing the above:

- Newton-Raphson
- Fisher scoring
- Iteratively reweighted least squares (IRLS)

Newton-Raphson

A classic algorithm for handling iterative solutions of nonlinear systems of equations is the *Newton-Raphson algorithm*:

- Start with an initial guess β_0 for the solution.
- Obtain a second guess by approximating the function to be maximized in a neighborhood of the initial guess by a second-degree polynomial and then finding the location of the polynomial's maximum value.
- Repeat until the discrepancy in successive evaluations of the objective function evaluate along the sequence of iterates is smaller than some convergence threshold.

The sequence of iterates that this algorithm generates converge to a solution $\hat{\beta}$ when the optimization function is suitable (full rank properly conditioned Fisher Information matrix) and/or the initial guess is good.

Let

- $U(\beta) = \nabla l(\beta)$ be the score function
- $H(\beta) = \nabla^2 l(\beta)$ denote the Hessian matrix.

At iteration k, consider the following second order Taylor series approximation of $l(\beta)$,

$$l(\beta) \approx l(\beta_k) + U(\beta_k)^T(\beta - \beta_k) + \frac{(\beta - \beta_k)^T H(\beta_k)(\beta - \beta_k)}{2}. \quad (1)$$

Now solving

$$U(\beta) \approx U(\beta_k) + H(\beta_k)(\beta - \beta_k) = 0$$

for β yields the next guess. That guess is

$$\beta_{k+1} = \beta_k - H(\beta_k)^{-1}U(\beta_k). \quad (2)$$

This algorithm is locally fast, exhibits quadratic convergence, provided that it converges.

Convergence is likely in identifiable models where $H(\beta_0)$ is positive definite.

However, the Newton-Raphson method can be quite sensitive to the choice of starting values β_0.

For many identifiable GLMs with full rank model matrices the Hessian is negative definite and the log likelihood is a strictly concave function. The maximum likelihood estimators of model parameters exist and are unique under quite general conditions.

Fisher scoring algorithm

The *Fisher scoring algorithm* is an alternative optimization method.

It resembles the Newton-Raphson algorithm, the distinction being with the Hessian matrix used in the Newton updates.

Fisher scoring uses the expected Fisher information matrix instead of the Hessian which is the observed Fisher information matrix.

We will let $\mathcal{H}$ be the expected information matrix so that $\mathcal{H}(\beta) = -\mathsf{E}\left\{\nabla^2 l(\beta)\right\}$.

The Newton update step for the Fisher scoring method is

$$\beta_{k+1} = \beta_k + \{\mathcal{H}(\beta_k)\}^{-1} U(\beta_k).$$

Notes

The chain rule allows us to write $\mathcal{H}(\beta) = M^T W(\beta) M$ where

$$W(\beta) = \left\{ \nabla_\theta^2 c(\theta^*) \right\} |_{\theta^* = M\beta}.$$

The estimated asymptotic covariance matrix $\mathcal{H}^{-1}$ of $\hat{\beta}$ occurs as a by-product of this algorithm as $\{\mathcal{H}(\beta_k)\}^{-1}$

For GLMs with canonical link (the entirety of this class), we have that the observed and expected information are the same.

For both Fisher scoring and Newton-Raphson, the score function $U(\beta)$ can be written as

$$U(\beta) = \nabla l(\beta) = M^T \left\{ Y - \nabla_\theta c(\theta^*)|_{\theta^*=M\beta} \right\}.$$

For noncanonical link models (which we see later), Fisher scoring has the advantages that it produces the asymptotic covariance matrix as a by-product

Iterratively reweighted least squares (IRLS)

Recall the linear regression set up where

$$Z = M\beta + \varepsilon$$

where V is the covariance matrix of ϵ.

The *weighted least-squares (WLS)* estimator of β is

$$\hat{\beta}_{\mathsf{WLS}} = \left(M^T V^{-1} M\right)^{-1} M^T V^{-1} Z.$$

We can write

$$\mathcal{H}(\beta) = M'W(\beta)M$$

and the score function as

$$U(\beta) = M'W(\beta)W^{-1}(\beta)(y - \mu(\beta)).$$

The "response" $Z(\beta)$ can be written as

$$Z(\beta) = M'\beta + W^{-1}(\beta)(y - \mu(\beta))$$

where $\mu(\beta)$ is the mean value parameter defined as a function of β.

The Newton update step of the Fisher scoring method can be written as

$$\mathcal{H}(\beta_k)\beta_{k+1} = \mathcal{H}(\beta_k)\beta_k + U(\beta_k). \tag{3}$$

The right hand side of (3) can be rewritten as

$$\mathcal{H}(\beta_k)\beta_k + U(\beta_k) = M^T W(\beta_k) z(\beta_k).$$

Thus,

$$(M^T W(\beta_k) M)\beta_{k+1} = M^T W(\beta_k) z(\beta_k)$$

The above equation has a solution of the form

$$\beta_{k+1} = (M^T W(\beta_k) M)^{-1} M^T W(\beta_k) z(\beta_k),$$

Sufficiency

A (possibly vector-valued) statistic is *sufficient* if the conditional distribution of the full data given this statistic does not depend on the parameter.

The interpretation is that the full data provides no information about the parameter that is not already provided by the sufficient statistic.

The principle of sufficiency follows: *all inference should depend on the data only through sufficient statistics.*

The Fisher-Neyman factorization criterion says that a statistic is sufficient if and only if the likelihood depends on the whole data only through that statistic.

Lemma

The canonical statistic vector of an exponential family is a sufficient statistic.

Sufficient dimension reduction is a whole field of study. However, the *OG* “sufficient dimension reduction” theory was about exponential families.

The so-called Pitman-Koopman-Darmois theorem (proved independently by three different persons in 1935 and 1936) says that

> *When we have IID sampling from a statistical model, all distributions in the model have the same support which does not depend on the parameter, and all distributions in the model are continuous, then there is a sufficient statistic whose dimension does not depend on the parameter if and only if the statistical model is an exponential family of distributions.*

Why is this theorem is written this way? Your humble instructor is still working on his writing.

Notes

This theorem was responsible for the interest in exponential families early in the twentieth century.

The condition of the Pitman-Koopman-Darmois theorem that the support does not depend on the parameter is essential.

The condition that the statistical model has to be continuous is ugly. Later theorems covered discrete distributions.

Sufficient dimension reduction for canonical linear submodels remains important.

Maximum entropy

Edwin Jaynes, a physicist, introduced the *maximum entropy formalism* that describes exponential families in terms of entropy.

Suppose we have a big exponential family (a *saturated model*) and are interested in submodels. The maximum entropy argument says the canonical linear submodels are the submodels that, *subject to constraining the means of their submodel canonical statistics, leave all other aspects of the data as random as possible*, where "as random as possible" means maximum entropy.

When we specify $\theta = M\beta$ we are, in effect, modeling only the the distribution of the submodel canonical statistic $t(y) = M'y$, leaving all other aspects of the distribution of y as random as possible given the control over the distribution of $t(y)$.

The relative entropy of a distribution with PMF f to a distribution with PMF m is defined to be

$$-\sum_{x \in S} f(x) \log\left(\frac{f(x)}{m(x)}\right),$$

where S is the support of the distribution with PMF m.

Suppose we know the value of some expectations

$$\mu_j = \mathrm{E}\left(t_j(X)\right) = \sum_{x \in S} t_j(x) f(x), \qquad j \in J.$$

Suppose we want f to maximize entropy subject to these constraints plus the constraints that f is nonnegative and sums to one:

$$\begin{aligned} \text{maximize } & -\sum_{x \in S} f(x) \log\left(\frac{f(x)}{m(x)}\right) \\ \text{subject to } & \sum_{x \in S} t_j(x) f(x) = \mu_j, \qquad j \in J \\ & \sum_{x \in S} f(x) = 1 \\ & f(x) \geq 0, \qquad x \in S. \end{aligned}$$

The maximizer of f is an exponential family!