STAT 528 - Advanced Regression Analysis II

Exponential family theory (part 3)

Daniel J. Eck
Department of Statistics
University of Illinois

Last time

- mean value parameters
- maximum likelihood estimation (MLE)
- asymptotics of MLE
- finite sample concentration of MLE

Learning Objectives Today

- generalized linear models (GLMs)
- different parameterizations
- motivation of logistic regression
- inference

Canonical linear submodels: intro to GLMs

We now motivate generalized linear models (GLMs) within the context of exponential theory.

A canonical affine submodel of an exponential family is a submodel having parameterization

$$\theta = a + M\beta$$

where:

- $\theta \in \mathbb{R}^n$ is the canonical parameter vector
- $\beta \in \mathbb{R}^p$ is the canonical parameter vector for the submodel
- $a \in \mathbb{R}^n$ is a known offset vector
- $M \in \mathbb{R}^{n \times p}$ is a known *model matrix*.

In most applications the offset vector is not used giving parameterization

$$\theta = M\beta,$$

in which case we say the submodel is a *canonical linear submodel*.

We will restrict attention to the canonical linear submodel in this class.

The canonical linear submodel log likelihood is given by

$$
\begin{aligned}
l(\theta) &= \langle y, \theta \rangle - c(\theta) \\
&= \langle y, M\beta \rangle - c(M\beta) \\
&= \langle M'y, \beta \rangle - c_\beta(\beta),
\end{aligned}
\tag{1}
$$

and we see that we again have an exponential family with

- canonical statistic $M'y$
- cumulant function $\beta \mapsto c_\beta(\beta) = c(M\beta)$
- submodel canonical parameter vector β

Stepping back a bit

In practice n denotes the sample size.

$\theta \in \mathbb{R}^n$ is an arbitrary unrestricted vector specifying one parameter for individual. This is a saturated model.

For a model to be useful, we need dimension reduction

$$\theta = M\beta.$$

In other words, $\theta \in \text{span}(M)$.

Full regular submodels

If the originally given full canonical parameter space was Θ, then the full submodel canonical parameter space is

$$B = \{\beta : M\beta \in \Theta\}.$$

Thus a canonical linear submodel gives us a new exponential family, with lower-dimensional canonical parameter and statistic.

The submodel exponential family is full regular if the original exponential family was full regular.

Parameterizations

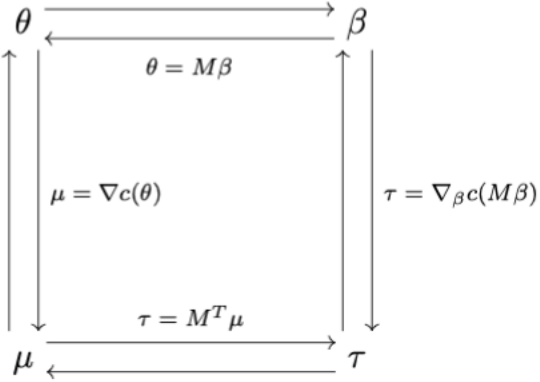
Now we have four parameters:

- the saturated model canonical and mean value parameters θ and μ
- the canonical linear submodel canonical and mean value parameters β and $\tau = M^T \mu$.

The observed equals expected property for the submodel is

$$\hat{\tau} = M^T \hat{\mu} = M^T y. \tag{2}$$

A depiction of the transformations necessary to change between parameterizations.



Note: $\mu \to M^T \mu$ is usually not one-to-one (when $n > p$ and M is full column rank).

Hence we cannot determine $\hat{\theta}$ and $\hat{\beta}$ from them either.

The only way to determine the MLEs is to maximize the log likelihood (1) to obtain $\hat{\beta}$ and then

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

GLMs and link functions

Recall that the saturated model canonical parameter vector θ is *linked* to the saturated model mean value parameter vector through the change-of-parameter mappings $g(\theta)$.

We can reparameterize $\theta = M\beta$ and write

$$\mu = \mathrm{E}_\theta(Y) = g(M\beta)$$

which implies that we can write

$$g^{-1}\left(\mathrm{E}_\theta(Y)\right) = M\beta.$$

Therefore, a linear function of the canonical submodel parameter vector is linked to the mean of the exponential family through the inverse change-of-parameter mapping g^{-1}.

This is the basis of exponential family generalized linear models with link function g^{-1}.

Note that most treatments of GLMs will present g^{-1} as the link function. Instead we motivated a change of parameters mapping from canonical to mean value parameters.

Example: logistic regression

Logistic regression is a different form of regression with binary outcomes.

Data is observed in independent pairs (y_i, x_i'), $i = 1,\ldots,n$ where the covariates are assumed to be known.

The idea is the probability of success changes with covariates in the following manner

$$p_i = P(Y_i|x_i) = f(x_i'\beta)$$

where $f : \mathbb{R} \to (0,1)$ is continuous and monotone.

We will study linear regression using $f = g$. Other choices of f will be considered later.

Let M have rows x_i', $i = 1, \ldots, n$.

Then $\theta_i = x_i'\beta$,

$$g(x_i'\beta) = \left(\frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)}\right) = p_i,$$

and

$$g^{-1}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = x_i'\beta.$$

We can write the log likelihood in canonical form in the saturated model parameterization starting from the log likelihood of independent Bernoulli trials:

$$\sum_{i=1}^{n} [y_i \log(p_i) - (1 - y_i) \log(1 - p_i)] = \langle y, \theta \rangle - c(\theta).$$

Alternatively, we can write the log likelihood in canonical form in the submodel parameterization starting from the log likelihood of independent Bernoulli trials:

$$\sum_{i=1}^{n} [y_i \log(p_i) - (1 - y_i) \log(p_i)] = \langle M' y, \beta \rangle - c_\beta(\beta)$$

Difference in symmantics

$$\nabla_\beta c(\theta) = \nabla_\beta c(M\beta) = M^T \left[\nabla_\theta c(\theta)|_{\theta=M\beta} \right]$$

vs

$$\nabla_\beta c_\beta(\beta) = M^T \left[\nabla_\theta c(\theta)|_{\theta=M\beta} \right].$$

Inference

We have a regular full exponential family submodel with log likelihood written in canonical form

$$l(\beta) = \langle M^T y, \beta \rangle - c(M\beta)$$

From here, we have that the asymptotic distribution of the MLE $\hat{\beta}$ takes the form

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \xrightarrow{d} N(0, \Sigma^{-1})$$

where $\Sigma = \mathrm{E}\left(-\nabla_\beta^2 l(\beta)\right)$ is the Fisher information matrix corresponding to the canonical linear submodel.

Wald inference for β

Let $\widehat{\Sigma}$ be estimated Σ using the MLE $\hat{\beta}$ in place of β. In particular, the jth element $\hat{\beta}_j$ of $\hat{\beta}$ is asymptotically normal with asymptotic variance

$$\widehat{\text{Var}}(\hat{\beta}_j) = jth \text{ diagonal element of } \widehat{\Sigma}.$$

The Wald Z statistic for testing $H_o : \beta_j = \beta_{jo}$ is

$$z_W = \frac{\hat{\beta}_j - \beta_{jo}}{\text{se}(\hat{\beta}_j)} \quad \overset{H_o}{\sim} \quad N(0,1)$$

where $\text{se}(\hat{\beta}_j) = \sqrt{\widehat{\text{Var}}(\hat{\beta}_j)}$.

We can construct $(1 - \alpha) \times 100\%$ Wald based confidence intervals of the form

$$\hat{\beta}_j \pm z_{\alpha/2}\mathsf{se}(\hat{\beta}_j)$$

for some error tolerance $0 < \alpha < 1$.

Inference for mean value parameters

One could consider a function $f : \beta \to \mu$ and then estimate the variance of $\hat{\mu}$ via the Delta method. One could then use Wald inference.

Alternatively, we could use the change of parameters map and plug-in:

$$\left[g(\hat{\beta} - z_{\alpha/2}\text{se}(\hat{\beta}), g(\hat{\beta} + z_{\alpha/2}\text{se}(\hat{\beta})\right]$$

- The former has the advantage that it is derived from principles.
- The latter has the advantage that it will always return valid values for mean value parameters.

Deviance, goodness of fit, and likelihood inference

To motivate the deviance of a statistical model we will rewrite the log likelihood as $l(\mu; y)$.

The unrestricted MLE of μ is y. Now consider a canonical submodel (GLM) of the form $\theta = M\beta$.

Let $\hat{\mu}$ be the MLE of μ restricted to an identifiable canonical submodel ($\hat{\mu} = \nabla c(\hat{\theta})$ where $\hat{\theta} = M\hat{\beta}$). It follows that

$$l(y; y) \geq l(\hat{\mu}; y).$$

The *deviance* of the GLM is

$$D(y; \hat{\mu}) = -2\left(l(\hat{\mu}; y) - l(y; y)\right).$$

The deviance statistic has approximate distribution

$$D(y; \hat{\mu}) \overset{H_o}{\sim} \chi^2_{\mathrm{df}}$$

where H_o is that the canonical submodel is correct and the alternative test is that the canonical submodel is incorrect but the saturated model is correct where

- $\mathrm{df} = n - p$
- n is the sample size
- p is the number of parameters in the canonical submodel.

We reject correctness of the canonical submodel when

$$D(y; \hat{\mu}) \; > \; \chi^2_{\mathsf{df}}(\alpha).$$

(Note that the χ^2 approximation can be poor.)

Wait, isn't this backwards?

A brief look at sufficiency

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

In other words, if $\theta = M\beta$ yields

$$D(y; \hat{\mu}) \; < \; \chi^2_{\mathsf{df}}(\alpha),$$

then we have sufficient dimension reduction from n to p.

Deviance testing for nested submodels

Let $\mathcal{M}_0$ and $\mathcal{M}_1$ both be canonical submodels.

We say that $\mathcal{M}_0$ is *nested* within $\mathcal{M}_1$ when every distribution in $\mathcal{M}_0$ is also in $\mathcal{M}_1$ but not vice-versa.

That is, μ is more restricted under $\mathcal{M}_0$ than under $\mathcal{M}_1$. Another way to look at it is that

$$M_1 = [M_0, A],$$

where

- M_0 is the model matrix for model $\mathcal{M}_0$
- M_1 is the model matrix for model $\mathcal{M}_1$
- A represents additional covariates collected for all individuals

Let

- $\hat{\mu}_0$ be the MLE of μ under $\mathcal{M}_0$
- $\hat{\mu}_1$ be the MLE of μ under $\mathcal{M}_1$.

We can use this framework for testing

$$H_0 : \mathcal{M}_0 \text{ true} \qquad H_a : \mathcal{M}_1 \text{ true, but not } \mathcal{M}_0$$

using the likelihood ratio χ^2 statistic given by

$$-2\left[l(\hat{\mu}_0; y) - l(\hat{\mu}_1; y)\right] \approx \chi^2_{\text{df}},$$

where

- $\text{df} = p_1 - p_0$
- p_1 is the number of parameters in $\mathcal{M}_1$
- p_0 is the number of parameters in $\mathcal{M}_0$.

(Note that the χ^2 approximation is often adequate here even it isn't adequate for the saturated model provided that $\mathcal{M}_1$ is not too close to saturated.)