# Aster Models for Life History Analysis with Lessons for Teaching Statistics

Charles J. Geyer

School of Statistics
University of Minnesota

February, 6, 2011

Joint work with Ruth Shaw, Stuart Wagenius, Julie Etterson, Helen Hangelbroek, Caroline Ridley, Robert Latta, Frank Shaw

# Code and Info

R contributed package `aster` on CRAN.

```
install.packages("aster")
library(aster)
```

Function `aster` fits models. Generic functions `summary`, `predict`, and `anova` work like those for linear and generalized linear models.

`http://www.stat.umn.edu/geyer/aster/` has links to papers and tech reports. All tech reports done with `Sweave` so everything is exactly reproducible.

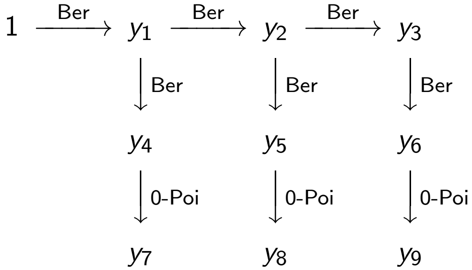
# An Aster Graph

$$
\begin{array}{ccccccc}
1 & \xrightarrow{\text{Ber}} & y_1 & \xrightarrow{\text{Ber}} & y_2 & \xrightarrow{\text{Ber}} & y_3 \\
 & & \big\downarrow \text{Ber} & & \big\downarrow \text{Ber} & & \big\downarrow \text{Ber} \\
 & & y_4 & & y_5 & & y_6 \\
 & & \big\downarrow \text{0-Poi} & & \big\downarrow \text{0-Poi} & & \big\downarrow \text{0-Poi} \\
 & & y_7 & & y_8 & & y_9
\end{array}
$$

$y_i$ are components of response vector for one individual (all individuals have isomorphic graphs). 1 is the constant 1.

Arrows indicate conditional distributions of variable at head of arrow (successor) given variable at tail of arrow (predecessor). Ber = Bernoulli, 0-Poi = zero-truncated Poisson.

# An Aster Graph (cont.)




Graph for *Echinacea angustifolia* example in Geyer, Wagenius and Shaw (*Biometrika*, 2007).

$y_1$, $y_2$, $y_3$ indicate survival in each of three years (2002–2004).

$y_4$, $y_5$, $y_6$ indicate flowering status ($1 =$ some flowers, $0 =$ no flowers) in corresponding years.

$y_7$, $y_8$, $y_9$ are flower counts in corresponding years.

# Abstract Aster Graph

Nodes (variables) have at most one predecessor, hence graph is specified by function $p$ that maps from set $J$ of non-initial nodes to set $N$ of all nodes. $y_{p(j)}$ is predecessor of $y_j$.

$y_j$ at initial nodes treated as constants. Then joint distribution factors as product of conditionals

$$f_\theta(y) = \prod_{j \in J} f_\theta(y_j \mid y_{p(j)})$$

because graph is not allowed to have loops. Log likelihood is

$$l(\theta) = \sum_{j \in J} \log f_\theta(y_j \mid y_{p(j)})$$

# Predecessor is Sample Size

In subgraph

$$y_{p(j)} \longrightarrow y_j$$

$y_j$ is sum of $y_{p(j)}$ independent and identically distributed random variables.

Define $\xi_j$ to be the mean of one of those variables, so

$$E(y_j \mid y_{p(j)}) = y_{p(j)} \xi_j$$

$\xi_j$ are components of *conditional mean value parameter* vector $\xi$.

# Predecessor is Sample Size (cont.)

Define

$$\mu_j = E(y_j)$$

components of *unconditional mean value parameter* vector $\mu$.

By iterated expectation theorem

$$\begin{aligned} E(y_j) &= E\{E(y_j \mid y_{p(j)})\} \\ &= E(y_{p(j)} \xi_j) \\ &= \xi_j E(y_{p(j)}) \end{aligned}$$

that is

$$\mu_j = \xi_j \mu_{p(j)}$$

# Predecessor is Sample Size (cont.)

$$
\begin{array}{ccccccc}
1 & \xrightarrow{\text{Ber}} & y_1 & \xrightarrow{\text{Ber}} & y_2 & \xrightarrow{\text{Ber}} & y_3 \\
 & & \downarrow{\scriptstyle \text{Ber}} & & \downarrow{\scriptstyle \text{Ber}} & & \downarrow{\scriptstyle \text{Ber}} \\
 & & y_4 & & y_5 & & y_6 \\
 & & \downarrow{\scriptstyle \text{0-Poi}} & & \downarrow{\scriptstyle \text{0-Poi}} & & \downarrow{\scriptstyle \text{0-Poi}} \\
 & & y_7 & & y_8 & & y_9
\end{array}
$$

$$\mu_1 = \xi_1$$
$$\mu_2 = \xi_2\xi_1$$
$$\mu_3 = \xi_3\xi_2\xi_1$$
$$\mu_4 = \xi_4\xi_1$$
$$\mu_5 = \xi_5\xi_2\xi_1$$

$$\mu_6 = \xi_6\xi_3\xi_2\xi_1$$
$$\mu_7 = \xi_7\xi_4\xi_1$$
$$\mu_8 = \xi_8\xi_5\xi_2\xi_1$$
$$\mu_9 = \xi_9\xi_6\xi_3\xi_2\xi_1$$

# Exponential Families of Distributions

An *exponential family* of distributions is a statistical model with log likelihood

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

when terms not containing the parameter have been dropped, and

$$\langle z, \theta \rangle = z^T \theta = \theta^T z$$

Statistic vector $z$ and parameter vector $\theta$ that give log likelihood of this form are called *canonical*. $c$ is called *cumulant function*.

Log likelihood for $z_1$, ..., $z_n$ independent and identically distributed is

$$\langle z_1 + \cdots + z_n, \theta \rangle - nc(\theta)$$

# Exponential Families and Predecessor is Sample Size

Each conditional distribution of $y_j$ given $y_{p(j)}$ is one-parameter exponential family having $y_j$ as the canonical statistic. In

$$l(\theta) = \sum_{j \in J} \log f_\theta(y_j \mid y_{p(j)})$$

$j$-th term of is

$$y_j \theta_j - y_{p(j)} c_j(\theta_j)$$

(compare with)

$$\langle z_1 + \cdots + z_n, \theta \rangle - n c(\theta)$$

Have subscripts on $c_j$ and $\theta_j$ because each arrow can have different exponential family and different parameter.

# Aster Model Log Likelihood

$$
\begin{aligned}
l(\theta) &= \sum_{j \in J} \bigl[ y_j \theta_j - y_{p(j)} c_j(\theta_j) \bigr] \\
&= \sum_{j \in J} y_j \left[ \theta_j - \sum_{\substack{k \in J \\ j = p(k)}} c_k(\theta_k) \right] - \sum_{\substack{k \in J \\ p(k) \notin J}} y_{p(k)} c_k(\theta_k)
\end{aligned}
$$

This is recognizable as log likelihood for joint exponential family.
Blue term is $j$-th component of joint canonical parameter vector.
Red term is cumulant function of joint family.

# Aster Model Log Likelihood (cont.)

$$l(\varphi) = \langle y, \varphi \rangle - c(\varphi)$$

where

$$\varphi_j = \theta_j - \sum_{\substack{k \in J \\ j = p(k)}} c_k(\theta_k), \qquad j \in J$$

and

$$c(\varphi) = \sum_{\substack{k \in J \\ p(k) \notin J}} y_{p(k)} c_k(\theta_k)$$

# Aster Transform

$\theta$ is the *conditional canonical parameter vector*.
$\varphi$ is the *unconditional canonical parameter vector*.

Map between them is invertible.

$$\theta_j = \varphi_j + \sum_{\substack{k \in J \\ j = p(k)}} c_k(\theta_k)$$

where $\theta_k$ on right-hand side have "already" been determined as function of $\varphi$. Use in any order that does successors before predecessors (always is one because graph has no loops).

# Exponential Family Canonical and Mean Value Parameters

By properties of exponential families

$$\xi_j = c_j'(\theta_j)$$
$$\mu = \nabla c(\varphi)$$

where prime denotes univariate derivative and $\nabla$ denotes multivariate derivative (vector of partial derivatives).

By properties of exponential families these changes of parameters are also invertible (although no closed-form expression in general, inversion equivalent to doing maximum likelihood).

# A Plethora of Parameters

Four different parameterizations $\mu$, $\xi$, $\theta$, and $\varphi$.

All are important. All play a role in some scientific arguments. Users have to understand all four.

But wait, there's more!

# Canonical Linear Submodels

In an exponential family, with canonical parameter $\varphi$, the change of parameter

$$\varphi = M\beta$$

where $M$ is a known matrix (model matrix or design matrix) gives a new exponential family because

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

and

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

Submodel canonical parameter vector is $\beta$.
Submodel canonical statistic vector is $M^T y$.
Submodel mean value parameter vector is $\tau = E(M^T y) = M^T \mu$.

# A Plethora of Parameters (cont.)

Six different parameterizations

| | | | |
|---|---|---|---|
| $\mu$ | saturated model | unconditional | mean value |
| $\xi$ | saturated model | conditional | mean value |
| $\varphi$ | saturated model | unconditional | canonical |
| $\theta$ | saturated model | conditional | canonical |
| $\beta$ | submodel | unconditional | canonical |
| $\tau$ | submodel | unconditional | mean value |

# Fisher Information

Fisher information for submodel canonical parameter vector $\beta$ is

$$I(\beta) = -\nabla^2 l(\beta) = M^T \nabla^2 c(M\beta) M$$

Computer can convert to any other parameterization. And compute derivatives for applying the delta method to transfer standard errors.

# Interpretation

In any exponential family, map between canonical and mean value parameter vectors is strictly multivariate monotone. Hence

$$
\begin{aligned}
\langle \mu_1 - \mu_2, \varphi_1 - \varphi_2 \rangle &> 0 \\
\langle \tau_1 - \tau_2, \beta_1 - \beta_2 \rangle &> 0 \\
\langle \xi_1 - \xi_2, \theta_1 - \theta_2 \rangle &> 0
\end{aligned}
$$

where $\mu_1$, $\varphi_1$, $\tau_1$, $\beta_1$, $\xi_1$, $\theta_1$ are different parameter vectors corresponding to the same aster model, ditto for $\mu_2$, $\varphi_2$, $\tau_2$, $\beta_2$, $\xi_2$, $\theta_2$, and the two models are distinct.

# Interpretation (cont.)

Multivariate monotonicity dumbed down.

If $\varphi_i$ increases, other components of $\varphi$ being held fixed, then $\mu_i$ increases (other components of $\mu$ also change).

If $\beta_i$ increases, other components of $\beta$ being held fixed, then $\tau_i$ increases (other components of $\tau$ also change).

If $\theta_i$ increases, then $\xi_i$ increases.

# Interpretation (cont.)

Maximum likelihood estimators (MLE) in exponential families have the "observed equals expected" property. If $\hat{\mu}$ and $\hat{\tau}$ are MLE, then

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

Submodel always fits perfectly those aspects of the data that are submodel canonical statistics (components of $M^T y$).

# Interpretation (cont.)

Exponential families have the maximum entropy property (Jaynes).

Each distribution in the family is as random as possible (maximizes entropy) subject to having a particular value of the mean value parameter vector $\tau = M^T \mu$.

If the components of the submodel canonical statistic vector $M^T y$ are scientifically interpretable, then the submodel is scientifically interpretable.

# Lessons

In teaching linear regression we usually introduce models with

$$y = M\beta + \text{error}$$

because students don't understand parametric statistic models.

Must be unlearned to understand generalized linear models.

# Lessons (cont.)

In teaching linear regression we usually introduce confidence intervals for the mean value parameter

$$\mu = M\beta$$

by not calling $\mu$ a "parameter" but "predicted values." Similarly for values of the regression function $x^T \beta$ for hypothetical data $x$.

This allows students to think of $\beta$ as "the" vector of parameters and never think about other parameters.

But it gives students wrong intuitions about confidence intervals. Confidence intervals are always for *parameters*. That there are confidence intervals for things you don't want to call parameters (even though they are) must also eventually be unlearned.

# Lessons (cont.)

In generalized linear models have three essential parameters.

Submodel parameter vector $\beta$.
Saturated model canonical parameter vector $\varphi = M\beta$.
Mean value parameter vector $\mu$.

Calling only $\beta$ "the" parameter vector and referring to $\varphi$ as the "linear predictor" and $\mu$ as the "predicted values" is is even more confusing in this context.

Confidence intervals for components of $\varphi$ and $\mu$ are often scientifically important.

# Lessons (cont.)

The map

$$\varphi = M\beta$$

between submodel and saturated model canonical parameter vectors is much woofed about.
(In linear regression, this is $\mu = M\beta$ because $\varphi = \mu$.)

In intro courses, where students don't know about matrices, this is presented as

$$\mu_i = x_{i1}\beta_1 + \cdots + x_{ip}\beta_p$$

or something of the sort.

Students are told this is how to interpret linear, generalized linear, and all regression-like models (including aster models).

# Lessons (cont.)

The dual map

$$\tau = M^T \mu$$

between saturated model and submodel mean value parameter vectors just as important, if not more important, to scientific interpretation, but ignored in teaching.