Coverpage

```
Document build date: Tue Aug 21 16:55:32 2012

Working directory :
      /home/millaco/work/git_projects/a4a/model-documentation

Current contents of .GlobalEnv:

     <empty>

Session information:

R version 2.15.1 (2012-06-22)
Platform: x86_64-pc-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C              
 [3] LC_TIME=en_GB.UTF-8        LC_COLLATE=en_US.UTF-8    
 [5] LC_MONETARY=en_GB.UTF-8    LC_MESSAGES=en_US.UTF-8   
 [7] LC_PAPER=C                 LC_NAME=C                 
 [9] LC_ADDRESS=C               LC_TELEPHONE=C            
[11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C       

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base     

other attached packages:
[1] xtable_1.7-0       lattice_0.20-6     RColorBrewer_1.0-5

loaded via a namespace (and not attached):
[1] grid_2.15.1  tools_2.15.1
```

Description of Catch at Age model

Colin P. Millar, Ernesto Jardim

1 Motivation

The goal is to describe an age based model that is robust and easy to use. Robustness here means robust parameter estimation in addition to being robust to (underlying) model complexity. We aim to provide robust parameter estimates by concentrating parameters out of the objective function where possible and using structured random effects to allow for complexity when the data support it. More on this in the coming sections, but first a brief description of the data that the model will have to fit to.

2 Data and model

The data are

C_{at} catch at age a and year t

S_{atk} abundance index for age a and year t from the kth survey or CPUE series, $k = 1, 2, \ldots$

The model is an age structure model where the number of fish in a given cohort N at the start of the following year is the number of fish that survived the perils of the current year. We assume that fish die through the year at a constant rate e^{-Z} (Z is positive), and that this rate is solely due to natural causes (M) and fishing (F) so that the total mortality rate is $Z = F + M$. This results in the model

$$N_{a+1,t+1} = N_{at} e^{-Z_{at}}$$

Abundance indices are observations of the relative abundance not of absolute abundance. This is because trawl surveys do not detect every fish but a fixed proportion Q. This proportion depends on age through length and means the index is proportional to abundance

$$S_{at} = Q_a N_{at}$$

If F and M are constant through the year catches arise as a fraction of those fish that died, and is written here as the familiar Baranov catch equation

$$\begin{aligned} C_{at} &= \frac{F_{at}}{Z_{at}} \left(N_{a+1,t+1} - N_{at} \right) \\ &= \frac{F_{at}}{Z_{at}} \left(1 - e^{-Z_{at}} \right) N_{at} \end{aligned}$$

These last two equations show that in there own way, catches and abundance indices are both observations of the numbers of fish in the population. Neither is sufficient to estimate the absolute abundances N but together they can be used to estimate both N and F. One way of doing this is using a statistical catch at age approach (see for example Myers and Cadigan, 1994) where much like in a regression model we find the values of F and N that give the best fit to our observations. The only specification we need to make is how our observations come about, or how are observations distributed about the model predictions. We assume that catches and indices are normally distributed on the log scale. This is for two reasons: 1) catches and indices are typically positive with variances that increase as the level increases in such a way that taking logs makes the variance constant. 2) the equations given above become easier to deal with on the log scale as we will show later. This leads to the following distributional assumptions

$$c_{at} \sim N\left(\log\left(\frac{F_{at}}{Z_{at}}\left(1-e^{-Z_{at}}\right)\right) + n_{at}, \quad \sigma_c^2 \right)$$

and

$$s_{at} \sim N\left(q_a + n_{at}, \quad \sigma_s^2 \right)$$

where in these equations we have written logs in lower case i.e. $c = \log C$.

2.1 Model fitting

The parameters that are estimated are log recruitment r_t, survey catchability q_a, the F parameters (MORE ON THESE LATER) and log stock numbers n_{a1} in the first year. The model is written in terms of these parameters,

$$\begin{aligned} c_{at} &= r_{t-a+1} - \sum_{i=1}^{a-1} Z_{a-i,t-i} + \log\left(\frac{F_{at}}{Z_{at}}\left(1-e^{-Z_{at}}\right)\right) + \epsilon_{at} \\ s_{at} &= q_a + r_{t-a+1} - \sum_{i=1}^{a-1} Z_{a-i,t-i} + \epsilon'_{at} \end{aligned}$$

where ϵ denotes the Gaussian observation error. There are some modifications required for the early cohorts as they use n_{a1} rather than recruits and for the plus groups, but these are trivial and not presented. It is possible to write these equations in matrix notation if we combine all catches into a single vector: $\boldsymbol{c} = (c_{11}, c_{21}, \ldots, c_{A_+1}, c_{12}, \ldots, c_{A_+2}, \ldots)^T$ and do similarly for the survey indices $\boldsymbol{s}$, it is also simpler if we define $\boldsymbol{r} = (n_{A1}, n_{A-1,1}, \ldots, n_{21}, r_1, \ldots)^T$ and combine the F model parameters into a single vector $\boldsymbol{f}$ then we can write the model as

$$\begin{aligned} \boldsymbol{c} &= \boldsymbol{X}_r\boldsymbol{r} + o_1(\boldsymbol{f}) + o_2(\boldsymbol{f}) + \boldsymbol{\epsilon} \\ \boldsymbol{s} &= \boldsymbol{M}\boldsymbol{X}_q\boldsymbol{q} + \boldsymbol{M}\boldsymbol{X}_r\boldsymbol{r} + \boldsymbol{M}o_1(\boldsymbol{f}) + \boldsymbol{\epsilon}' \end{aligned}$$

where the functions o_1 and o_2 are nonlinear (vector) functions of the F model parameters and the $\boldsymbol{X}$ matrices are various design matrices based on the full set of ages and years (to be described later). The $\boldsymbol{M}$ matrix maps the correct age and year in the survey to that in the full set of ages and years. These equations can be further combined by stacking the equations

$$\begin{pmatrix}\boldsymbol{c}\\ \boldsymbol{s}\end{pmatrix} = \begin{pmatrix}\boldsymbol{0} & \boldsymbol{X}_r \\ \boldsymbol{M}\boldsymbol{X}_q & \boldsymbol{M}\boldsymbol{X}_r\end{pmatrix}\begin{pmatrix}\boldsymbol{q}\\ \boldsymbol{r}\end{pmatrix} + \begin{pmatrix}\boldsymbol{I} & \boldsymbol{I} \\ \boldsymbol{M} & \boldsymbol{0}\end{pmatrix}\begin{pmatrix}o_1(\boldsymbol{f})\\ o_2(\boldsymbol{f})\end{pmatrix} + \begin{pmatrix}\boldsymbol{\epsilon}\\ \boldsymbol{\epsilon}'\end{pmatrix}$$

so that the model is of the form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + o(\boldsymbol{f}) + \boldsymbol{\epsilon}$$

where

$$\boldsymbol{\epsilon} \sim N\left(\boldsymbol{0}, \boldsymbol{W}^{-1}\right) \quad \text{where} \quad \boldsymbol{W}^{-1} = \begin{pmatrix}\sigma_c^2\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_s^2\boldsymbol{I}\end{pmatrix} = \sigma_c^2\begin{pmatrix}\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \frac{\sigma_s^2}{\sigma_c^2}\boldsymbol{I}\end{pmatrix}$$

In other words, this statistical catch at age model can be written as a linear model with an offset due to nonlinear functions of the F model parameters. We use this to reduce the parameters in the fitting process by concentrating the likelihood. This can be done by inserting the maximum likelihood estimates of β conditional on $\boldsymbol{f}$ and $\boldsymbol{W}$ into the likelihood. The maximum likelihood estimate of β conditional on the other parameters is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{y} - o(\boldsymbol{f}))$$

If we decide that the surveys and catches have the same observation variance then we can also estimate the obervation variance conditionally, in this case

$$\begin{aligned}
\hat{\boldsymbol{\beta}}(\boldsymbol{f}) &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{y} - o(\boldsymbol{f}))\\
\hat{\boldsymbol{y}}(\boldsymbol{f}) &= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{y} - o(\boldsymbol{f})) + o(\boldsymbol{f})\\
&= \boldsymbol{H}(\boldsymbol{y} - o(\boldsymbol{f})) + o(\boldsymbol{f})\\
&= \boldsymbol{H}\boldsymbol{y} + (1 - \boldsymbol{H})o(\boldsymbol{f})
\end{aligned}$$

and the conditional (unbiased) estimate of σ^2 is

$$\hat{\sigma^2}(\boldsymbol{f}) = \frac{1}{n-p}\log\Big((\hat{\boldsymbol{y}}(\boldsymbol{f}) - \boldsymbol{y})^T(\hat{\boldsymbol{y}}(\boldsymbol{f}) - \boldsymbol{y})\Big)$$

where n is the length of $\boldsymbol{y}$ and p is the number of unique parameters in β. The concentrated log likelihood for $\boldsymbol{f}$ is then

$$\begin{aligned}
l(\boldsymbol{f}) &\propto -\frac{n}{2}\log\Big((\hat{\boldsymbol{y}}(\boldsymbol{f}) - \boldsymbol{y})^T(\hat{\boldsymbol{y}}(\boldsymbol{f}) - \boldsymbol{y})\Big)\\
&= -\frac{n}{2}\log\left(\Big|\Big|(1 - \boldsymbol{H})\boldsymbol{y} + (1 - \boldsymbol{H})o(\boldsymbol{f})\Big|\Big|\right)
\end{aligned}$$

Since $1 - \boldsymbol{H}$ and $(1 - \boldsymbol{H})\boldsymbol{y}$ only depend on the data these can be calculated outside of an iterative optimization procedure. It is straightforward to give different weights to the survey and catch components without increased computation (this is the same as assuming you know the ratio $\frac{\sigma_c^2}{\sigma_s^2}$). However, if you want to estimate both variances then the estimate of β is

$$\begin{aligned}
\hat{\boldsymbol{\beta}}(\boldsymbol{f}) &= (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{y} - o(\boldsymbol{f}))\\
\hat{\boldsymbol{y}}(\boldsymbol{f}) &= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{y} - o(\boldsymbol{f})) + o(\boldsymbol{f})
\end{aligned}$$

and as $\boldsymbol{W}$ is not known before hand the inverse $(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}$ must be computed at every iteration. The concentrated log likelihood in this case is

$$\begin{aligned}
l(\boldsymbol{f}) &= \frac{1}{2}\log|\boldsymbol{W}| - \frac{1}{2}\Big((\hat{\boldsymbol{y}}(\boldsymbol{f}) - \boldsymbol{y})^T\boldsymbol{W}(\hat{\boldsymbol{y}}(\boldsymbol{f}) - \boldsymbol{y})\Big)\\
&= -\frac{n_c}{2}\log\sigma_c^2 - \frac{n_s}{2}\log\sigma_s^2 - \frac{1}{2}\Big((\hat{\boldsymbol{y}}(\boldsymbol{f}) - \boldsymbol{y})^T\boldsymbol{W}(\hat{\boldsymbol{y}}(\boldsymbol{f}) - \boldsymbol{y})\Big)
\end{aligned}$$

3 Extending the model (brief)

First potential models for q_a. Linear forms can be included my simply changing the design matrices. By linear forms i mean spline smoothers with fixed degrees of freedom. An interesting addition would be to use penalised splines or better (i think) structured random effects (GMRFs i.e. 1st order and 2nd order random walks, see for example Rue and Held, 2005). Structured random effect models using GMRFs require a Bayesian approach. If the variance of these is assumed known then these random effects can be integrated out directly, but if the variance is to be estimated it must be done in the outside iteration along with the observation error and the F parameters.

The model for F has not been considered so far. Options for this are a simple separable model, a model with several separable periods all these models can be expressed as linear models on the log link. Within the same framework is simple then to include smoothers (splines) with fixed degrees of freedom. More interesting but perhaps out with the scope of this project are structured random effect models for F. These include seasonal models (treating the number of ages as the season length) and correlated random walks.

4 Summary

At its simplest the model is a non-linear fixed effects regression, fitting the F parameters to the data but this requires that the survey variances are known relative to the catch variance. This model can include splines, or random effects with known parameters (variances, degrees of freedom, autocorrelation) in the q and r models. The reason this is being considered is that computational speed is important and allowing some fixed variability may reduce bias in model outputs by allowing some flexibility, it is acknowledged that such assumptions make statistical testing a bit dodgy.

The next level of complexity is where we want to estimate the survey variance. This means we have to recalculate the hat matrix at every iteration which involves a matrix inversion. Since the matrix inversion is being done already, including structured random effects for q and recruitment into this only adds the variance parameters to the objective function.

If a stock recruit function were to be added, recruits could not be integrated out and would have to be estimated in the objective function.

If more complicated F models were used, such as structured random effects (random walks, correlated random walks, seasonal models) the number of parameters would increase in the objective function. Therefore a model with few parameters is one with a highly parametrised F model and no SRR relationship.

5 Implementation (sketch)

We require two functions that return an objective

A inputs are F at age and observation error and arguments are the data and the hat matrix

B inputs are F at age and any variance parameters taking as arguments the design matrix H, the weight matrix W and the structural prior matrix (not mentioned yet but lets call it Q)

The full objective function is then

1. take input parameters (F pars, variances, recruitments (if SRR model being used))
2. convert F pars into F at age
3. calculate objective value using one of the two functions A or B above
4. add on SRR density and prior densities for variances if necessary

References

Myers, R. A. and N. G. Cadigan (1994). The statistical analysis of catch-at-age data with correlated errors. *NAFO N2422*, 1–11.

Rue, H. and L. Held (2005). *Gaussian Markov Random Fields: Theory and Applications.* Chapman and Hall/CRC.