Identifiability (a nasty issue)

There are PIMs that specify models to *MARK* that cannot be "identified." That is, the data do not permit the estimation of some parameters. This is an inherent lack of information that keeps some parameters from being estimated.

The issue is like the estimation of the regression model

$$E(y) = \beta_0 + \beta_1(x)$$

when the sample size $n = 1$ (only one sample of y and x). Here, the model parameters cannot be identified because there is an infinite number of lines that can be drawn through a point. Of course, if one had a sample size of 2, then a unique like would be defined and the 2 parameters could be estimated (or "identified").

In the fish tagging or bird band recovery models, identifiability of survival would not be an issue if fish were tagged in only a single year AND reporting probability was a constant across years. Consider the bass data, for example,

2000 30 70 114 43 15 1728,

corresponding to the cell probabilities under model $\{S, r\}$;

$$R_1 \quad (1\text{-}S)r \quad S(1\text{-}S)r \quad SS(1\text{-}S)r \quad SSS(1\text{-}S)r \quad SSSS(1\text{-}S)r \quad .$$

One can see that S can be estimated (not an MLE) as

$$\hat{S} = m_{14}/m_{13} = SSS(1\text{-}S)r \Big/ SS(1\text{-}S)r = S\ .$$

Note, the r in the numerator and denominator cancel, as do the terms $(1–S)$ and 2 of the S, leaving $\hat{S} = S$. Numerically $\hat{S} = 43/114 = 0.377$. So, under this simple model, the constant survival probability can be estimated, if r is also constant across time.

Under models where r is allowed to vary by year or age, then identifiability is lost, unless more than one cohort is tagged and released.

A common headache in model $\{S_t\ r_t\}$ is the lack of identifiability of the terms shown below in bold:

Number Tagged	Matrix of Probabilities for m_{ij} E(m_{ij}/R_i)				
R_1	$(1\text{-}S_1)r_1$	$S_1(1\text{-}S_2)r_2$	$S_1S_2(1\text{-}S_3)r_3$	$S_1S_2S_3(1\text{-}S_4)r_4$	$S_1S_2S_3S_4\mathbf{(1\text{-}S_5)r_5}$
R_2		$(1\text{-}S_2)r_2$	$S_2(1\text{-}S_3)r_3$	$S_2S_3(1\text{-}S_4)r_4$	$S_2S_3S_4\mathbf{(1\text{-}S_5)r_5}$
R_3			$(1\text{-}S_3)r_3$	$S_3(1\text{-}S_4)r_4$	$S_3S_4\mathbf{(1\text{-}S_5)r_5}$
R_4				$(1\text{-}S_4)r_4$	$S_4\mathbf{(1\text{-}S_5)r_5}$
R_5					$\mathbf{(1\text{-}S_5)r_5}$

In this case, only the product $\mathbf{(1\text{-}S_5)r_5}$ is identifiable, but not the separate terms. Thus, this model has 4 survival probabilities, 4 reporting probabilities and one product term that can be identified under model $\{S_t\ r_t\}$. Total, K = 9 (not 10, as you might think/want).

This subject will haunt us continually and more insights will be provided (the concepts of sufficient and minimal sufficient statistics and their dimensionality). Program *MARK* has clever ways to help understand this issue, but is not perfect for complicated or ill-conditioned models.

More on Identifiability and Related Issues

Any model for tag recovery data is based on interpretable parameters, especially of the type S and r (or S and f) explicitly appearing in the model structure for E(m_{ij} | R_i). However, just because a parameter appears in the model does not mean that parameter can in fact be estimated from data. Most parameters in the model are estimable, but not all; it depends on the model.

The idea of parameters not being estimable is illustrated by trying to estimate S_1 from one year of tag recoveries, m_{11}:

$$\frac{\text{E}(m_{11})}{R_1} = (1 - S_1)r_1.$$

You cannot do it; you only have an estimate of the product, $(1 - S_1)r_1$ and there is no way to separately estimate S_1 and/or r_1.

For all models we know (or can know) what parameters are estimable and hence we know the number, K, needed for QAIC$_c$ (or likelihood ratio tests). (For some models this information is embedded in the help file of *MARK*). You do not need to know K to fit models. *MARK* tries to determine K by numerical methods, it does not always succeed. So there are times when K needs to be input to *MARK* to get QAIC$_c$ computed correctly. You also need to beware of interpreting numerical results for non-estimable parameters as meaning anything. You can tell such cases by the estimated standard error: it will be either huge, or paradoxically, trivially small (near zero).

There is another practical problem that arises in model fitting: point estimates that are on a boundary (i.e., and $\hat{S} = 1$, or $r = 1$). This can lead to *MARK* computing the wrong *K* for the model. Also, when this occurs, it suggests the model is too general, hence not the best one to use. When a parameter estimate is “pegged" on a boundary its estimated standard error will generally be quite wrong (too small), and this event (estimate on a boundary) can cause the estimated standard errors of other estimates to be wrong. You need to look at fitted models to be sure no such anomalies have occurred and to see if *MARK* has *K* correct, so QAIC_c is correct. The issue of the correct *K* is a difficult one for us to provide advice about. However, anomalous point estimates and weird standard errors are indicators of either basic parameter non estimability, or may just reflect sparse (poor) data or a bad fitting model.

An example of a problem with sparse data that produced an estimate on a boundary is the below. First, the input data were

```
/*  Release  recovery  data  for  RELEASES >=  711  MM  (28
INCHES)  long,  data  for  Chesapeake  Bay,  from  Cynthia
Goshorn via Dave Smith.  Years are 1987 to 1996  */
recovery matrix group=1;
 1    0    2    0    1    2    1    0    0    0;
      6    8    7   14    6    1    3    0    0;
           9   17   17    6    4    3    5    2;
               23   16   11    5    2    4    0;
                    47   24   20    4    9    3;
                         44   28   18   16    7;
                              58   44   40   11;
                                   52   42   22;
                                        61   29;
                                             92;
29 129 221 304 396 438 628 545 529 862;
```

The first year of releases and that cohort of recoveries are too small to be worth including in the analysis. However, dropping that year does not solve the problems with analysis of these data under model {*S*(t), *r*(t)}. Output under this model follows.

model={*S*(t), *r*(t)}

I	S(I)		Standard Error	95% Confidence Interval Lower	Upper
1	0.6882106		0.2549591	0.1769578	0.9577357
2	0.9597202	←	0.0162191	0.9128110	0.9818921
3	0.9477962	←	0.0117022	0.9194888	0.9665133
4	0.5873970		0.0721886	0.4426089	0.7184953
5	0.5735394		0.0754826	0.4234658	0.7111909
6	0.6190184		0.0745679	0.4664646	0.7512166
7	0.6884013		0.0816988	0.5115350	0.8233431
8	0.7018095		0.0960572	0.4890719	0.8526551
9	0.4699632		0.0792827	0.3221038	0.6232899
10	0.0000000		0.0000000	0.0000000	0.0000000
	r(I)				
11	0.1105959		0.1346352	0.0084308	0.6452122
12	1.0000000	←	0.5302730E-05	0.9999896	1.0000104
13	1.0000000	←	0.4885705E-05	0.9999904	1.0000096
14	0.1755289		0.0378107	0.1131433	0.2621452
15	0.2866250		0.0637381	0.1790485	0.4253479
16	0.2762157		0.0667091	0.1655747	0.4232850
17	0.3217343		0.0938853	0.1695157	0.5243401
18	0.3133693		0.1132253	0.1399389	0.5614302
19	0.2263497		0.0466889	0.1478443	0.3303798
20	0.1067285		0.0105167	0.0877991	0.1291613

" ← " indicates problem estimates

What went wrong? Insight can be gained by looking at a parameterization from Brownie et al. (1985) where

$$f_j = (1 - S_j)/r_j \ .$$

Under this model parameterization, the unrestricted MLEs of S_2 and S_3 are > 1

model={S(t), f(t)}

I	S(I)		Standard Error	95% Confidence Interval Lower	Upper
1	0.5931043		0.2271318	0.1872679	0.9021626
2	1.0797344	←	0.1820650	0.7228870	1.4365818
3	1.1707343	←	0.1903254	0.7976965	1.5437720
4	0.5099335		0.0772790	0.3620754	0.6560738
5	0.5735387		0.0754824	0.4234656	0.7111900
6	0.6270057		0.0755834	0.4715178	0.7600288
7	0.6796323		0.0807166	0.5063700	0.8143737
8	0.7018086		0.0960566	0.4890726	0.8526536
9	0.4699636		0.0792828	0.3221041	0.6232904
	f(I)				
10	0.0344828		0.0338830	0.0048356	0.2079196
11	0.0410397		0.0164941	0.0184737	0.0886799
12	0.0501508		0.0117315	0.0315548	0.0788141
13	0.0628728		0.0104667	0.0452223	0.0867861
14	0.1222342		0.0137126	0.0977977	0.1517497
15	0.1052331		0.0119675	0.0839769	0.1310996
16	0.0989749		0.0099066	0.0811819	0.1201578
17	0.0934437		0.0099059	0.0757538	0.1147516
18	0.1199738		0.0125178	0.0975225	0.1467532
19	0.1067285		0.0105167	0.0877991	0.1291613

This makes it clear that two estimates are out of range and this creates problems. We prefer to treat such cases as a diagnostic that a model with too many parameters have been used for the analysis of the data.