Data Dredging – Two Examples

An Example:

The perils (and rewards) of data dredging can be illustrated using the data on lake trout tagged in Cayuga Lake, NY (from Robson and Youngs, 1971). The data on angler recoveries are given below in the format for program MARK:

```
/* Cayuga lake trout data, Robson and Youngs 1971 */
 recovery matrix group=1;
    72 44  8  9  4  4  1  1  1  0;
       74 30 20  7  4  2  1  0  0;
          54 48 13 23  5  4  2  0;
             74 24 16  7  3  1  1;
                48 40  5  5  2  5;
                   31 10  6  3  2;
                      38 30  6  2;
                         19  6  6;
                            13 14;
                               21;
   1048 844 989 971 863 465 845 360 625 760;
```

The analyst has no particular objectives in mind and decides to run a general model and examine the estimates of annual survival probabilities. He runs model $\{S_t, f_t\}$ using program MARK with the following results (the first 9 estimates are the S_i, while the last 10 estimates are the f_i):

			95% Confidence Interval	
Parameter	Estimate	Standard Error	Lower	Upper
1	**0.4201790**	**0.0579119**	**0.3126108**	**0.5359043**
2	**0.4754611**	**0.0635581**	**0.3548625**	**0.5989914**
3	**0.7178074**	**0.0884068**	**0.5195336**	**0.8568106**
4	**0.4809078**	**0.0674984**	**0.3528886**	**0.6114839**
5	**0.6318942**	**0.1068580**	**0.4109425**	**0.8085744**
6	**0.4502940**	**0.0884398**	**0.2890974**	**0.6226487**
7	**0.5478015**	**0.1197612**	**0.3195595**	**0.7575658**
8	**0.6957785**	**0.1998632**	**0.2643676**	**0.9357124**
9	**0.7328569**	**0.2311426**	**0.2133440**	**0.9652168**
10	0.0687023	0.0078136	0.0548805	0.0856895
11	0.0918754	0.0090832	0.0755554	0.1112962
12	0.0575123	0.0064096	0.0461641	0.0714412
13	0.0712518	0.0070869	0.0585511	0.0864544
14	0.0510052	0.0061238	0.0402545	0.0644342
15	0.0713281	0.0101249	0.0538574	0.0939037
16	0.0427691	0.0060010	0.0324356	0.0562036
17	0.0560535	0.0104174	0.0388040	0.0803300
18	0.0229500	0.0050918	0.0148281	0.0353609
19	0.0276315	0.0059468	0.0180820	0.0420086

This model has an **AIC value of 7482.28 and *K* = 19.**

In studying the MLEs he notices that year 6 was a very hard winter with many days of sub-zero temperatures and thick surface ice on the lake. He hypothesizes that this might explain the low estimated survival in year 6 (i.e., $\hat{S}_6 = 0.45$). He also sees that $\hat{S}_1 = 0.42$ (even lower) and is not sure of an explanation for this low estimate, but believes the winter in that year might also have been quite severe. Showing the effect of hard winters on lake trout survival probabilities would be very important and surely angler management implications would be appropriate. He decides to run a model with survival in years 1 and 6 (the two years where estimated survival probability was lowest) as one parameter and survival in all the other years as a second parameter; thus model $\{S_{1\&6}, S_{all\ others}, f_t\}$. The results are (the f_i are shown for completeness – sorry to show so many significant digits):

			95% Confidence Interval		
Parameter	Estimate	Standard Error	Lower	Upper	
1	**0.4198689**	**0.0446044**	**0.3357641**	**0.5089004**	**2 hard winters**
2	**0.5870801**	**0.0229384**	**0.5415156**	**0.6312004**	**all other years**
3	0.0684718	0.0077818	0.0547054	0.0853894	
4	0.0912461	0.0086460	0.0756565	0.1096671	
5	0.0531361	0.0055127	0.0433118	0.0650374	
6	0.0754894	0.0063072	0.0640198	0.0888189	
7	0.0473447	0.0050653	0.0383493	0.0583221	
8	0.0713293	0.0074610	0.0580194	0.0874092	
9	0.0441293	0.0056064	0.0343565	0.0565192	
10	0.0546559	0.0068872	0.0426251	0.0698345	
11	0.0248779	0.0043508	0.0176345	0.0349905	
12	0.0325624	0.0046868	0.0245283	0.0431118	

This model has **AIC = 7475.05 and *K* = 12**. This is clearly a better model *for these data*; Δ = 7482.28 – 7475.05 = 7.23. Note the confidence intervals for the two estimated survival probabilities ($S_{1\&6}$, vs. $S_{all\ others}$). This person could submit a manuscript showing the association of hard winters on trout survival. The likelihood of acceptance for publication might be quite high.

There are several problems with this *post hoc* approach. First, is the high risk that the result is, in fact, spurious. If data were available for, say, 1980-91 the effect of the so-called hard winter might not be supported. The effect seen in this data set might be unique, and not a part of the process of interest. Second, the biology was put in place after studying the estimates from a general model. Third, the "hard winter" in year 1 was barely justified and, indeed, it had the lowest estimated survival, i.e., the idea of a "hard winter" was suggested to

the analyst by the data, and not a question conceived before inspection of the data and estimates. One might ask if several of the high survival years were associated with years where the winter was quite mild (the other side of the coin). Lastly, notice there is almost a "reward" for data dredging – one almost always finds something "significant." Journal editors are rarely leery of observational studies where *post hoc* data dredging has lead to the "finding" of some pattern or association. There is little or no "penalty" for data dredging! Perhaps that explains why data dredging has become so common.

A Second Example:

Ken Burnham made several analyses of the Cayuga Lake trout data. He first considered 5 models based on *a priori* interests. These are summarized below, in the usual notation:

Model	AICc	Delta AICc	AICc Weight	K
{S(T) f(t)}	7478.406	0.00	0.62853	12
{S(.) f(t)}	7480.016	1.61	0.28101	11
{S(t) f(t)}	**7482.283**	**3.88**	**0.09046**	**19**
{S(t) f(.)}	7536.240	57.83	0.00000	10
{S(.) f(.)}	7561.245	82.84	0.00000	2

Here, it is unclear what the data might support in terms of an inference. A linear trend in survival probability is the best approximation; however, the model where survival is constant across years is a close contender (Akaike weight = 0.28). Even the model with unspecified annual variation in the survival probabilities is not a terrible model (Akaike weight = 0.09). At this point, the analyst might present the table above in a manuscript, but be perhaps hesitant to say much about patterns in survival probability, given year-specific recovery rates in each case. Perhaps the honest inference concerning survival probabilities is that we do not know much about its variation in this case.

His colleague enters the room and looks at the MLEs from model $\{S_t, f_t\}$ (above) and notes that these split out nicely at 0.5. That is, he sees that about half (4) of the estimates are < 0.5, whereas the other half (5, actually) are > 0.5. He suggests a model of *bad* and *good* years, with 0.5 serving as the cut-off point. Without much thinking, they agree to run this *post hoc* model, with the following results:

Model	AICc	Delta AICc	AICc Weight	K
{S(good>0.5),S(bad <0.5),f(t)}	7470.745	0.00	0.96663	12
{S(T) f(t)}	7478.406	7.66	0.02097	12
{S(.) f(t)}	7480.016	9.27	0.00938	11
{S(t) f(t)}	7482.283	11.54	0.00302	19
{S(t) f(.)}	7536.240	65.49	0.00000	10
{S(.) f(.)}	7561.245	90.50	0.00000	2

Not only is this new model the best, but it is the best by a wide margin (Δ = 7.66, and w = 0.97 for the *post hoc* model vs. 0.02 for the 2^{nd} ranked model). The MLEs from this new model are similarly “interesting":

			95% Confidence Interval		
Parameter	Estimate	Standard Error	Lower	Upper	
1	**0.4604916**	**0.0277539**	**0.4067731**	**0.5151440**	**bad**
2	**0.6616833**	**0.0415927**	**0.5760646**	**0.7378784**	**good**
3	0.0684511	0.0077764	0.0546939	0.0853564	
4	0.0892965	0.0081924	0.0744883	0.1067091	
5	0.0577793	0.0061317	0.0468703	0.0710380	
6	0.0744287	0.0062405	0.0630844	0.0876223	
7	0.0532202	0.0056075	0.0432415	0.0653443	
8	0.0711083	0.0072514	0.0581422	0.0866998	
9	0.0422063	0.0052457	0.0330413	0.0537722	
10	0.0485277	0.0063758	0.0374520	0.0626655	
11	0.0217379	0.0039959	0.0151421	0.0311159	
12	0.0284038	0.0044751	0.0208313	0.0386205	

Data dredging can certainly be "effective" (note the confidence intervals above). However, as a basis for *valid* inference, data dredging is dangerous. In this case, one has strong evidence that survival in *bad* years is less than that in *good* years (by definition, if nothing else!).

More emphasis should be placed on *a priori* thinking and modeling, without blatant data dredging and so much *post hoc* inference. Their are many cases where the *post hoc* results are not misleading (not spurious) and exploration of the data is a useful thing to do.

However, more caution is needed in accepting exploratory results as if they were somewhat confirmatory. Often, we suspect the researcher does not know, or care, that *post hoc* results are often spurious. Journal editors surely care, but do not seem to act in a manner that will tend to minimize the publication of spurious results.

Acknowledgments:

Dr. Steve Cherry (Montana State University) contributed material for this note.

Literature Cited:

Robson, D. S., and W. D. Youngs. 1971. Statistical analysis of reported tag-recaptures in the harvest from an exploited population. Biometrics Unit, Cornell University, BU-369M. 15pp.