Choosing number of stages of multistage model for cancer modeling: SOP for contractor and IRIS analysts

Definitions in this memo:

1. *Order of the multistage model* is the highest power term in the multistage equation. For example, $P(x; \beta) = \beta_0 + \beta_1 * x + \beta_2 * x^2$ is of order 2 because 2 is the highest power in multistage equation. Also note that number of estimated parameters is always one more than the order. In this example, there are 3 estimated parameters.
2. *Dose groups* are all the groups in the experiment, including the control group. For example, an experiment with doses 0, .1, 1, 10 has 4 dose groups
3. *Estimate on the boundary* is the estimate of one of the β parameters that is equal to 0. Note that if the parameter's true value is assumed to be 0 and is fixed at zero, this is not a 'parameter on the boundary' situation
4. *Adequate model fit* (as detailed in Section 2.3.5 of the BMD Technical Guidance document) for the multistage model is as follows: goodness-of-fit p-value > 0.05, scaled residuals < |2|, and good low-dose fit.

Instructions:

1. Fit all orders of the multistage model up to two less than the number of dose groups. In other words, if there are k dose groups, fit up to model order k-2. Fitting a k-2 order model to a dataset will leave 1 degree of freedom (df) with which to calculate the goodness-of-fit p-value;
    a. If all parameter (γ, $\beta_{1, \ldots,} \beta_{k-2}$) estimates are positive, then:
        i. Use the AIC to select the best-fitting model if at least one of the models provides an adequate fit to the data;
        ii. If none of the models adequately fit the data, try a k-1 order model. If the k-1 order model fits adequately, use that model. Adequate fit could be determined by looking at each dose group's scaled residuals and the visual fit of the model to the data. Flag the matter (i.e., using order k-1) for the NCEA Statistics Workgroup (SWG) and the Assessment Manager (AM).
    b. Otherwise (i.e., if any parameter is estimated to be zero and is thus at a boundary), use the following procedure (2):
2. Examine fits of order 1 and 2 (linear and quadratic, respectively). Examine the linear (parameters: γ (background), β_1) and quadratic model (parameters: γ (background), β_1, β_2) for adequate fit;
    a. If neither model fits adequately, refer the matter to SWG and AM;
    b. If only one of the models fits adequately, use that model;
    c. If both models fit adequately:

i. Use the model with the lowest AIC if all of the parameters (γ , β_1,and β_2) are positive;

ii. Otherwise, use the model with the lower BMDL (more health protective). If BMD/BMDL ratio is larger than 3, flag this for the AM and SWG

Discussion

Model selection is discussed in USEPA (2012), Section 2.3.9, page 39, "Selecting the model to use for POD computation". See also Sections 2.3.5 through 2.3.8.

This SOP differs from current practice in the BMDS training, which is to fit the highest order possible (k-1 = one less than number of dose groups). The BMDS training cancer example (http://www.epa.gov/ncea/bmds/bmds_training/application/appl.htm#Example) considers both AIC and a LRT, and states "Under the recommendations of the benchmark dose guidance, the more parsimonious first-degree model would be generally preferred.", but does not identify a single criterion (AIC, LRT, parsimony). This change in practice is motivated by recent work by Nitcheva et al. (2007), discussed below, and by the fact that when k parameter estimates (k = number of dose groups) are non-zero the goodness-of-fit test cannot be made. This practice is expected to be appropriate for a large percentage of data sets (Nitcheva et al. 2007), but not for all, in which case higher-order multistage models and other types of models may be tried (see 2, above).

This SOP differs from a practice used by some statisticians, which is to use likelihood ratio tests (LRT) to conduct a statistical hypothesis test (SHT) comparing model orders. This is applied by stepping up (comparing 1st to 2nd order, 2nd to 3rd, etc.) until the test for the next higher order is not significant. It is based upon the table of log-likelihoods and degrees of freedom reported by BMDS. The method assumes that twice the LR statistic is asymptotically distributed as Chi-square.

Technical Background

AIC assumes that that no parameters are on the boundary (i.e., equal to zero) and thus AIC is not reliable for evaluating fit of multistage models, for which some parameters often will be on the boundary (much more often than for other models). See Claeskens and Hjort (2008).

A paper by Nitcheva et al. (2007) used formal statistical testing (accounting for parameters on the boundary) to show that for 91 IRIS datasets (having from 3 to 7 groups), about 80% are best fitted by a linear model and about 20% by the quadratic model. For none of the 14 datasets where a cubic model was fitted did the cubic model give sufficient improvement in fit (as determined by LRT). The authors noted "… the use of the higher-order term adds little or no significance to the quality of the fit."

The approach of using standard LRT asymptotics to select a multistage model is flawed. The null hypothesis for each test for the next order ***k*** can be characterized as Ho: $\beta_k = 0$ versus Ha: $\beta_k > 0$ (with other model parameters being nuisance parameters). There is a statistical objection to this one-sided test procedure (see Mohlenberghs and Verbeke 2007, Nitcheva et al. 2007, Kopylev and Sinha 2011, Kopylev 2012). A correct test requires use of specialized asymptotic distributions that depend on parameters of the model. Even in the simplest case, 1st vs. 2nd order, Ho: $\beta_2 = 0$ versus Ha: $\beta_2 > 0$, with Background $\gamma > 0$ and $\beta_1 > 0$, the mixture is half $\chi^2(1)$ and half zero. The distribution is more complicated when Background is at or near zero, and when more coefficients are added (and any of these are at or near zero). In such cases, the distributions can only be computed by simulation for each case. It is thus feasible, but not practicable, to conduct such tests rigorously for orders higher than three, and not even for 2nd order models when background estimate is zero.

Alternative procedure

It would be feasible and practicable to conduct a valid test as done by Nitcheva et al. (2007) to compare model orders 1 and 2. For routine use, this would require some expense to have expert contractor personnel write and test code for the procedure (we could also attempt to obtain code used by Nitcheva et al. to decrease effort). However, we do not believe that routine use of this procedure for a formal statistical hypothesis test would change outcomes or add much value, compared to the simple procedure described above.

If for some dataset there is special interest in higher-order models, these methods could be implemented.

References

Claeskens, G. and N.L. Hjort. 2008. Model Selection and Model Averaging. Cambridge U.P.

Kopylev, L. 2012. Constrained Parameters in Applications: Review of Issues and Approaches. ISRN Biomathematics Volume 2012, Article ID 872956, 6 pages. doi:10.5402/2012/872956

Kopylev, L., and B. Sinha. 2011. On the asymptotic distribution of likelihood ratio test when parameters lie on the boundary. Sankhya(B) **73**:20-41

G. Molenberghs and G. Verbeke. 2007. Likelihood ratio, score, and Wald tests in a constrained parameter space. American Statistician 61:22–27.

Nitcheva, D.K., W.W.Piegorsch and R.W.West. 2007. On use of the multistage dose–response model for assessing laboratory animal carcinogenicity. Regulatory Toxicology and Pharmacology 48:135-147.

USEPA. 2012. Benchmark Dose Technical Guidance. Publication No. EPA/100/R-12/001

Example 1

Data from Acrylonitrile assessment draft modeling appendix, for
Male Sprague-Dawley Rats, Zymbal Gland Tumors, Administered Dose Metric (Quast 2002)

Dose	Incidence	N
0.00	3	73
3.42	4	45
8.53	3	47
21.20	16	44

Model order[a]	**Goodness of fit**				**$BMD_{10\%}$ (mg/kg-d)**	**$BMDL_{10\%}$ (mg/kg-d)**
	***p*-value**	**Scaled residuals**	**AIC**	**Coefficients ***		
Two df=2	0.407	-0.255 0.934 -0.896 0.248	137.814	γ = 0.0474 β_1 = 0 β_2 = 0.0008363	11.2	6.25
One df=2	0.148	0.292 0.156 -1.645 1.001	140.414	γ = 0.0348 β_1 = 0.0148069	7.12	4.79
* From cancer models fitted in BMDS 2.40 by J.Fox, 14 May 2014						

BMDS cancer model counts only the non-zero parameters (including background and betas), so the Deviance test and Chi-square Goodness-of-Fit test both the 2nd and 1st degree models have 4-2=2 degrees of freedom (df).

In this case, the β_1 parameter for the 2nd degree multistage model has been estimated on the boundary (i.e., equal to zero). So, following the direction of Step 2 above, we would consider only the linear (i.e., 1st degree) and quadratic (i.e., 2nd degree models), in this case, the only models actually used. Both the 1st and 2nd degree models provide adequate fit to the data (Step 2.c). As the β_1 parameter was estimated as zero for the 2nd degree model, according to Step 2.c.ii, the model with the lowest BMDL would be selected as the best model. In this example, that is the 1st degree model, with a BMD = 7.12 and a BMDL = 4.79.

Example 2

If we replace the Incidence (3) at zero dose with a zero (0) as follows:

Dose	Incidence	N
0.00	0	73
3.42	4	45
8.53	3	47
21.20	16	44

The modeling results are:

Model order[a]	**Goodness of fit**				**$BMD_{10\%}$ (mg/kg-d)**	**$BMDL_{10\%}$ (mg/kg-d)**
	***p*-value**	**Scaled residuals**	**AIC**	**Coefficients ***		
Two df=2	0.174	0 1.346 -1.266 0.289	114.428	$\gamma = 0$ $\beta_1 = 0.0128276$ $\beta_2 = 0.00032934$	6.97	4.38
One df=3	0.319	0 0.841 -1.526 0.69	112.912	$\gamma = 0$ $\beta_1 = 0.0178669$	5.90	4.26
* From cancer models fitted in BMDS 2.40 by J.Fox, 15 May 2014						

In this example, the Deviance test and Chi-square Goodness-of-Fit test have 3 df (4-1=3) for the 1st degree model and 2 df (4-2=2) for the 2nd degree model.

In this case, the γ parameter has been estimated as zero for both the 1st and 2nd degree models. As this parameter has been estimated on the boundary, all other β coefficients for both models are estimated as non-zero. Following Step 2.c.i above, the model with lowest BMDL should be chosen. In this example, the 1st degree model would be selected as the best model, with a BMD = 5.90 and a BMDL = 4.26).

Example 3

Data for Bromate, for testicular mesothelioma tumors male F344/N rats
From Nitcheva et al. (2007), Table 1

Dose	Incidence	N
0	0	71
1.1	4	73
6.1	5	73
12.9	11	71
28.7	31	67

The modeling results are:

Model order[a]	Goodness of fit				$BMD_{10\%}$ (mg/kg-d)	$BMDL_{10\%}$ (mg/kg-d)
	***p*-value**	**Scaled residuals**	**AIC**	**Coefficients ***		
Three df=2	0.1602	-1.007 1.584 -0.321 -0.189 0.047	231.261	γ = 0.0140719 β_1 = 0.01066274 β_2 = 0 β_3 =1.25521E-005	9.04	5.32
Two df=2	0.1350	-1.095 1.596 -0.270 -0.389 0.187	231.771	γ =0.016603 β_1 = 0.00768476 β_2 = 0.00044062	9.03	5.01
One df=3	0.0717	-0.456 1.870 -1.047 -1.118 0.977	231.647	γ = 0.00291998 β_1 = 0.0179367	5.87	4.62
* From cancer models fitted in BMDS 2.40 by J. Allen Davis, 15 May 2014						

In this example, the Deviance test and Chi-square Goodness-of-Fit test have 3 df (5-2=3) for the 1st degree model, and 2 df (5-3=2) for the 2nd and 3rd degree models.

The β_2 parameter for the 3rd degree model is estimated on the boundary. Therefore, according to Steps 2, only the 1st and 2nd degree models should be considered further. Considering the fit statistics, it appears that both models fit the data adequately. Therefore, according to Step 2.c.i, the model with lowest AIC would be chosen as the best-fitting model. In this case, that is the 1st degree model, with a BMD = 5.87, and a BMDL = 4.62.