Goodness-of-fit in Product Multinomial Models

At first glance, goodness-of-fit (GOF) testing seems easy in, for example, a set of band recovery data. One could use the standard Pearson GOF test, where each term is of the form,

$$\frac{\textbf{(observed – expected)}^2}{\textbf{expected}} \quad \textbf{or} \quad \frac{\textbf{(O - E)}^2}{\textbf{E}}$$

Then, sum up these quantities over rows (j) and columns (i) of a recovery matrix as,

$$\mathbf{T} = \sum_i \sum_j \left(\frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \right),$$

where O_{ij} is the observed number of birds recovered for year j from birds banded in year i and $\hat{E}_{ij}$ are the corresponding estimated expected values. Under reasonable conditions, the test statistic **T** is asymptotically distributed as a χ^2 variable.

Consider the set of bird banding data from Brownie et al. (1985:2),

i	N_i	m_{ij} (or O_{ij})		
		j=1	2	3
1	1603	127	44	37
2	1595		62	76
3	1157			82

The observed data are merely the frequencies of bands reported (e.g., O_{11} = 127, O_{13} = 37, O_{33} = 82). The expected values (E_{ij}) for each cell under a particular model are not known, but these can be easily *estimated*, $\hat{E}_{ij}$, after the MLEs for a particular model are computed. It is important to note that the expectations vary by model. For example, under model $\{S, r\}$ of the Seber dead recoveries data type of Program MARK, the estimated expected values are,

$$\hat{E}_{11} = N_1(1 - \hat{S})\hat{r} = 1603 \cdot 0.0621 = 99.6$$

$$\hat{E}_{13} = N_1\hat{S}\hat{S}(1 - \hat{S})\hat{r} = 36.8, \text{ and}$$

$$\hat{E}_{33} = N_3(1 - \hat{S})\hat{r} = 71.9,$$

where N_i is the number banded in "year" i and $\hat{S}$ and $\hat{r}$ are the MLEs under model $\{S, r\}$.

Full expectations for two commonly-used models are shown below, for review.

Model $\{S_t, r_t\}$

Number Tagged	**Expected Number of Tags Reported (dead)**		
	j=1	2	3
N_1	$N_1(1-S_1)r_1$	$N_1S_1(1-S_2)r_2$	$N_1S_1S_2(1-S_3)r_3$
N_2		$N_2(1-S_2)r_2$	$N_2S_2(1-S_3)r_3$
N_3			$N_3(1-S_3)r_3$

Model $\{S, r\}$

Number Tagged	**Expected Number of Tags Reported (dead)**		
	j=1	2	3
N_1	$N_1(1-S)r$	$N_1S(1-S)r$	$N_1SS(1-S)r$
N_2		$N_2(1-S)r$	$N_2S(1-S)r$
N_3			$N_3(1-S)r$

These data could have been shown with the number *never recovered*:

i	N_i	m_{ij}			
		j=1	**2**	**3**	**never**
1	**1603**	**127**	**44**	**37**	**1395**
2	**1595**		**62**	**76**	**1457**
3	**1157**			**82**	**1075**

In fully parameterized models, such as $\{S_t, r_t\}$, there is no GOF information in the last cell "never recovered" but in other models, such as model $\{S, r\}$, GOF information is contained in the final cells. The number never recovered from banding in year 2 is 1457 and its expectation under Model $\{S_t, r_t\}$ is:

$$N_2 - \left(N_2(1-\hat{S}_2)\hat{r}_2 + N_2\hat{S}_2(1-\hat{S}_3)\hat{r}_3\right) = 1457.$$

The estimated expected number never recovered varies by model; e.g., for Model $\{S, r\}$ the expectation for year 2 is

$$N_2 - \left(N_2(1-\hat{S})r + N_2\hat{S}(1-\hat{S})r\right) = 1435.7$$

and is computed given the parameter estimates $\hat{S} = 0.6082$ and $\hat{r} = 0.1585$, under model $\{S, r\}$. Thus, computation of the estimated expected values ($\hat{E}_{ij}$) is relatively straightforward. Now, a small example, using the bird banding data from Brownie et al. (1985):
The data --

i	N_i	m_{ij} (or O_{ij})			
		j=1	2	3	never
1	1603	127	44	37	1395
2	1595		62	76	1457
3	1157			82	1075

The estimated expectations under model $\{S, r\}$--

i	N_i	$\hat{E}_{ij}$			
		j=1	2	3	never
1	1603	99.6	60.5	36.8	1406.1
2	1595		99.1	60.2	1435.7
3	1157			71.9	1085.1

The matrix of chi-squared contributions --

i	N_i	$\frac{(O - \hat{E})^2}{\hat{E}}$			
		j=1	2	3	never
1	1603	7.57	4.49	0.00	0.09
2	1595		13.86	4.16	0.31
3	1157			1.43	0.09

Each cell value is roughly a χ^2 variable with 1 df, thus, values > 3.84 might be viewed with suspicion (i. e., some evidence of lack of fit in that cell). Study of the matrix of chi-squared values provides evidence of lack of fit in several of the cells. Summing up the 9 values above, we get the test statistic **T** = 31.01 with 4 df and $P = 0.0000019$. One must conclude a substantial lack of fit of these data to the assumptions of model $\{S, r\}$. Under the more general model $\{S_t, r_t\}$ we get **T** = 1.80, with 1 df, $P = 0.1793$. This model seems to fit OK, but there is only a single degree of freedom to assess fit.

IMPORTANT PRACTICAL PROBLEMS

Practical problems arise in assessing model fit in all forms of capture-recapture and band recovery models. The problem involves cells where the expectation is small (e.g., < 0.1). Consider the banding data (above) in hypothetical year $j = 9$, where a single bird was

observed. Lets say, for example, that the estimated expectation for this cell is 0.1 (that is, $\hat{E}_{19} = 0.1$). Then, the chi-squared value for this single cell is

$$\frac{(O_9 - \hat{E}_9)^2}{\hat{E}_9} = \frac{(1 - 0.1)^2}{0.1} = \mathbf{8.1.}$$

This is "highly significant" but due to only the recovery of a single bird! A general rule of thumb for chi-squared issues is to be sure the expectations are > about 2. This is a useful rule. Lets look at one more example where a single bird was recovered, and let the corresponding expected value be 0.03; then we have,

$$\frac{(O - E)^2}{E} = \frac{(1 - 0.03)^2}{0.03} = \mathbf{31.36.}$$

Clearly, the test statistic in this case is hardly distributed as χ^2. A data set might have a dozen cells, all fitting nearly perfectly (i.e., $O_{ij} \sim \hat{E}_{ij}$) except one, where the chi-squared contribution was 31.36. The overall test statistic would strongly suggest a significant lack of fit, however, this evidence would be based on the recovery of a single bird! The distribution of the test will not provide interpretable information when expectations for some cells are small, even when only a single animal is recovered. Unfortunately, this is a very common case in the analysis of capture-recapture and band recovery data. There is, at present, no general solution to this issue. [Of course, small expectations are not an issue when no animals were recovered for that cell.]

In the band recovery models, an *ad hoc* pooling approach seems to be roughly sufficient. That is, starting from the right-most cells, pool over cell expectations to the left until the sum is > 2. Early software such as ESTIMATE and BROWNIE take this simple approach and it has been quite useful. Program RELEASE uses many 2x2 tables to assess GOF in capture-recapture models; numbers in these tables was often a pooling over cells. In these cases, a program option allows the use of Fisher's exact test for 2x2 tables and this was often useful. Still, cells where expected values are small represent a general problem.

OTHER APPROACHES TO GOF ASSESSMENT

While Pearson's GOF test is simple, commonly-used, and appealing, a likelihood-based alternative is useful and slightly superior. This procedure is based on

$$G^2 = 2\sum_i \sum_j O_{ij} \log_e \left(\frac{O_{ij}}{\hat{E}_{ij}} \right),$$

where $\mathbf{G^2}$ is also approximately (i.e., asymptotically) chi-square distributed. This is often termed the G^2 test or statistic. Program MARK makes extensive use of the G^2 statistic, because this statistic is the same as the deviance defined below. Both Pearson's and the G^2 test are special cases of the power-divergence statistic (see Read and Cressie, 1988, Springer-Verlag).

There are other general approaches to the GOF issue. One class of tests was first derived by Robson and Youngs (1971); see full discussion of this issue in Burnham et al. (1987:64-77). This type of test is often a 2xC contingency table and termed TEST2 in Burnham et al. (1987). It is a very useful test, applicable to both the band recovery models and the capture-recapture models, and is also asymptotically χ^2 distributed. Programs MARK and RELEASE provide the user with this general GOF test and TEST3, which is useful only in the capture-recapture models. While TEST2 and TEST3 are very useful and somewhat robust to pooling, there are still potential problems where cell expectations are small.

MORE ON SATURATED MODELS

Consider a model similar to model $\{S_t, f_t\}$ but where all the parameters are also specific to the cohort (i.e., year):

Number Tagged	**Matrix of Cell Probabilities for m_{ij}**			**E(m_{ij}/N_i)**
N_1	$(1-S_{11})r_{11}$	$S_{11}(1-S_{12})r_{12}$	$S_{11}S_{12}(1-S_{13})r_{13}$	
N_2		$(1-S_{22})r_{22}$	$S_{22}(1-S_{23})r_{23}$	
N_3			$(1-S_{33})r_{33}$	

The first subscript indexes the released cohort (all the parameters are cohort-specific), whereas the second subscript indexes the year of recovery. Here, 6 recovery probabilities and 3 survival probabilities appear in the model structure. None of the survival probabilities are estimable; only the initial recovery rates $[(1-S_{ii})r_{ii}]$ for each cohort are estimable. The MLE for each cell is merely the number observed in that cell divided by the number banded (i.e., m_{ij}/N_i). This is a fully-saturated model where there are as many unknown parameters as there are cells. It might just as well be expressed as

N_1	θ_{11}	θ_{12}	θ_{13}
N_2		θ_{21}	θ_{23}
N_3			θ_{33}

This makes it more clear that $\hat{\theta}_{13}$ = 37/1603 = 0.0231 (using the data in the example above). Of course, knowing that $\hat{\theta}_{13}$ = 0.0231 is of no biological interest as it is a confounding of $\{S_{11}S_{12}(1-S_{13})r_{13}\}$. Note, the saturated model always fits the data perfectly (by definition

and design). The concept of a saturated model is useful in computing **Deviance**. The deviance of model *j* is defined as

$$\textbf{Deviance} = -2\log_e(\mathcal{L}_j(\hat{\theta})) - -2\log_e(\mathcal{L}_{sat}(\hat{\theta})) .$$

If sample size is large (i.e., there are no cells with small expectations), then the deviance is asymptotically χ^2 with df = # cells in saturated model – # of estimable parameters in model *j*. Deviance is a type of GOF test, if sample size is large, and is exactly the same value as the G^2 test defined above. Note, also, that **T,** G^2, and **Deviance** are all asymptotically χ^2 distributed, but might vary substantially with sample sizes often seen in practice.

PROGRAM MARK BOOTSTRAP APPROACH

The goodness-of-fit of the global model can be evaluated in 3 ways: assuming that the deviance for the model is chi-square distributed and computing a goodness-of-fit test from this statistic, using Program RELEASE (for live recapture data only) to compute the goodness-of-fit tests provided by that program, and using the parametric bootstrap procedure provided in MARK.

The first approach is generally not valid because the assumption of the deviance being chi-square distributed is seldom met. This approach only seems reasonable for very large band recovery datasets, and this approach has never been reasonable for live recapture data because of the large number of possible histories that an animal may encounter. Use of Program RELEASE is reasonable, but usually lacks statistical power to detect lack of fit because of the amount of pooling required to compute chi-square distributed test statistics. For these reasons, the bootstrap procedure was implemented in MARK, available from the Results Browser menu..

With the bootstrap procedure, the estimates of the model being evaluated for goodness of fit are used to generate data, i.e., a parametric bootstrap. The simulated data exactly meet the assumptions of the model, i.e., no over-dispersion is included, animals are totally independent, and no violations of model assumptions are included. Data are simulated based on the number of animals released at each occasion. For each release, a simulated encounter history is constructed. As an example, consider a live recapture data set with 3 occasions (2 survival intervals) and an animal first released at time 1. The animal starts off with an encounter history of 100, because it was released on occasion 1. Does the animal survive the interval from the release occasion until the next recapture occasion? The probability of survival is ϕ_1, provided from the estimates obtained with the original data. A uniform random number in the interval (0, 1) is generated, and compared to the estimate of ϕ_1. If the random number is less than or equal to ϕ_1, the animal is considered to have survived the interval. If the random value is greater than ϕ_1, the animal has died. Thus, the encounter history would be complete, and would be 100. Suppose instead that the animal survives the first interval. Then, is it recaptured on the second occasion? Again, a new random number is generated, and compared to the capture probability p_2 from the parameter estimates of the model being tested. If the random value is less than p_2, the animal is considered to be captured, and the encounter history would become 110. If not captured, the encounter history would remain 100. Next,

whether the animal survives the second survival interval is determined, again by comparing a new random value with ϕ_2. If the animal dies, the current encounter history is complete, and would be either 100 or 110. If the animal lives, then a new random value is used to determine if the animal is recaptured on occasion 3 with probability p_3. If recaptured, the third occasion in the encounter history is given a 1. If not recaptured, the third occasion is left with a zero value.

Once the encounter history is complete, it is saved for input to the numerical estimation procedure. Once encounter histories have been generated for all the animals released, the numerical estimation procedure is run to compute the deviance and its degrees of freedom. These values along with $\hat{c}$ (= deviance / df) are saved to a simulation output file. The entire process is repeated for the number of simulations requested.

When the requested number of simulations is completed, the user can access the bootstrap simulations results database to evaluate the goodness of fit of the model that was simulated. First, the deviances of the simulated data can be ranked (sorted into ascending order), and the relative rank of the deviance from the original data determined. Suppose that the deviance of the original model was 101.01, whereas the largest deviance from 1000 simulations was only 90.90. Then you can conclude that the probability of observing a value as large as 101.01 was less than 1/1000. As another example, suppose the 801th simulated deviance in the sorted deviance file is 100.90, and the 802nd value was 101.50. Then, you would conclude that your observed deviance was reasonably likely to be observed, with probability of 198/1000 (because 198 of the simulated values exceeded the observed value).

A similar procedure can be used to evaluate the observed c-hat by comparing its rank to the simulated values of c-hat. Typically, conclusions using c-hat and deviance are about the same, but different results may be obtained with sparse data sets where the degrees of freedom associated with the deviance vary a lot across the simulations.

The bootstrap simulations can also be used to estimate the over-dispersion parameter, c. Two approaches are possible, based on the deviance directly, and on $\hat{c}$. For the approach based on deviance, the deviance estimate from the original data is divided by the mean of the simulated deviances to compute c-hat for the data. The logic is that the mean of the simulated deviances represents the expected value of the deviance under the null model of no violations of assumptions (i.e., perfect fit of the model to the data). Thus, $\hat{c}$ = observed deviance divided by expected deviance provides a measure of the amount of over-dispersion in the original data.

The second approach to estimating c for the original data is to divide the observed value of $\hat{c}$ from the original data by the mean of the simulated values of $\hat{c}$ from the bootstraps. Again, the mean of the simulated values provides an estimate of the expected value of c-hat under the assumption of perfect fit of the model to the data.

We're not sure of the benefits/disadvantages of the 2 procedures, and normally recommend using observed deviance divided by the mean of the bootstrap deviances because this approach does not rely on estimating the number of parameters, so is much faster.

Bootstrap Options allows you to specify that you are only interested in the deviance, and not $\hat{c}$, from the bootstrap simulations. Generally, results are about the same, but can be different when the degrees of freedom of the deviance varies a lot across the bootstrap simulations (caused by a small number of releases).

One of the current limitations of the bootstrap goodness-of-fit procedure is that individual covariates are not allowed.

LITERATURE CITED

Brownie, C., D. R. Anderson, K. P. Burnham, and D. S. Robson. 1985. Statistical inference from band recovery data a handbook. 2 Ed. U. S. Fish and Wildlife Service, Resource Publication 156. Washington, D. C., USA. 305pp.

Burnham, K. P., Anderson, D. R., White, G. C., Brownie, C., and Pollock, K. H. (1987). *Design and analysis methods for fish survival experiments based on release-recapture.* American Fisheries Society, Monograph **5.** 437pp.

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.

Seber, G. A. F. 1970. Estimating time-specific survival and reporting rates for adult birds from band returns. Biometrika 57:313-318.