

Model-Based Inference in the Life Sciences: A primer on evidence

by
David R. Anderson

ANSWERS TO SELETED EXERCISES

CHAPTER 2

Question 2. One worry might be that the algorithm did not converge to the maximum. This could be because the log-likelihood was very flat near the maximum point or that sparse data cause a lack of estimability. A second concern might be that mixture distributions are not in the exponential family and the log-likelihood function might have more than a single mode. Both issues might be helped by using different starting positions for the parameters. A more sophisticated approach would be to use simulated annealing.

Question 3. This is an easy question, but tricky if one is not thinking about the Principle of Parsimony! Instead of estimating 8 parameters, only 2 are required for the models under consideration, thus precision is often increased. In addition, one has a smoothed estimate of the functional response and this can be used for prediction outside the range of the data. This is another case where modeling has several rewards.

Question 4. This is a great question to ask of senior faculty members in various departments. See if they can point you to a model that exactly represents full reality. Proving that a model was the exact true model seems very difficult.

Question 6. This question should be easy if you understood the issue behind Question 3 (above). The parsimonious model had 6-10 parameters, whereas the full binomial model had 52 parameters that had to be estimated. The 52 estimates had high variances and this hampers interpretation.

CHAPTER 3

Question 1. I have no good answer to Cox's statement. I know of no literature where one objective (e.g., prediction) calls for one selection method whereas another objective (e.g., explanation) calls for another. Theory seems lacking; perhaps I am unaware of some aspect of the literature on this? There is a "focused Information Criterion" that tries to focus on a subset of the model parameters; I doubt if this was the subject of Cox's concern.

Question 2. To me, this ecotone issue suggests that PhD students ought to take a course or two well outside of their discipline (e.g., econometrics?).

Question 3. The classic example is the variance for samples from normally distributed populations. The MLE is $\sum(x_i - \hat{\mu})^2/n$ and the bias-adjusted form is $\sum(x_i - \hat{\mu})^2/(n-1)$ -- this is, of course, the usual least squares estimate. Adjusting estimators for bias is standard practice in mathematical statistics. However, in some sense, there is often too much concern for small sample bias. Bayesians do not say much about such bias in estimators. I consider the issue of such bias to be over-rated.

Question 9. This is a clear case where AICc should be used as an estimate of expected K-L information.

Question 10. This student is asking a good question. A good answer starts with a clear explanation of the Principle of Parsimony and what it has to say about under- and over-fitting and how bias decreases as more parameters are introduced; however, the uncertainty is increasing as there new parameters must be estimated. Also related her is the concept of tapering effect size and that these effects can be sequentially revealed as more data (and information) are available.

Question 11. The moral might be just to always use AICc. If sample size is small or if the number of parameters is large, then AICc better estimates expected K-L information.

Question 12. This is a hard question and I doubt if many experts in the model selection arena would want to stick their necks out very far on this issue. My thought would be that both are correct and have a sound underlying rationale. I would venture that the results would be very similar. In particular, if model averaging were done I would be the results would be virtually identical. If pressed into a choice I might prefer TIC just in case none of the models were of value.

Question 13. My approach would be to examine 2 models: one with a dummy variable for gender and another model without this variable. All other variables would be the same across the 2 models. Then look at the evidence ratio for the two models. Other approaches could rest on the model likelihoods or model probabilities. These approaches seem superior to the traditional t -test on the β value for the binary gender variable. Here the null hypothesis would be that $\beta \equiv 0$, and the alternative hypothesis is that $\beta \neq 0$. Then one selects an α value, hopes that the sampling distribution is well approximated by the t distribution, computes a P -value, and makes a decision concerning "statistical significance."

A related note deals with asymptotics in null hypothesis testing. Sample size can be infinite (or lets say $n =$ a billion trillion) and the investigator will still see an error α percent of the time! This is inconsistent with reasoning; why would the method be incorrect 5% of the time even with huge sample size. Clearly α should be a decreasing function of sample size by this is not the case in null hypothesis testing.

CHAPTER 4

Question 1. Model Selection Statistics for the Finch Bill Data

Model	$\log(\mathcal{L})$	K	AIC	Delta	w_i
1	-66.21	2	136.42	25.48	0.0000019
2	-57.77	5	125.54	14.60	0.0004488
3	-59.43	6	130.86	19.92	0.0000314
4	-60.98	6	133.96	23.02	0.0000067
5	-49.47	6	110.94	0	0.6643147
6	-49.47	7	112.94	2.00	0.2443877

7 -49.46 8 114.92 3.98 0.0908088

Selected Parameter Estimates

Model	$\hat{\beta}_1$	$\hat{se}(\hat{\beta}_1)$	$\hat{\beta}_2$	$\hat{se}(\hat{\beta}_2)$	$\hat{\beta}_3$	$\hat{se}(\hat{\beta}_3)$
1	--		--		--	
2	--		--		--	
3	0.07	0.44	--		--	
4	0.14	0.29	--		--	
5	1.63	0.30	--		--	
6	<u>0.21</u>	<u>0.38</u>	1.59	0.29	--	
7	<u>0.20</u>	<u>0.37</u>	1.55	0.34	<u>0.03</u>	<u>0.47</u>

- In Model 3 the β modifies T (a linear time trend)
- In Model 4 the β modifies X_1 (human disturbance)
- In Model 5 the β modifies X_2 (wet/dry years)
- In Model 6 the β s modify both X_1 and X_2
- In Model 7 the β s modify both X_1 and X_2 and the interaction

These β_i values are parameters in the submodels for the mixture coefficient π [not to be confused with the parameters α and β in the gamma distribution].

- a. The best hypothesis corresponds to model 5. It is "best" in the sense of smallest K-L information loss or that it is the hypothesis/model that is closest to full reality. It is too easy to say it has the smallest AICc value (it does) or it has a delta value of 0 (it does). While true, these are proxy statements; the real issue reflects back to K-L information. This is a fine point but worth remembering when writing a scientific manuscript.
- b. A clean approach to answering this question is the evidence ratio between models 1 and 2; this is about 236 and seems fairly convincing. Other evidence can be examined such as the fact that all the models of bimodality are better than the unimodal model.
- c. This calls for a value judgement. I believe this could be considered strong evidence against unimodality. People used to think "1 in 20" was "significant"; surely 1 in 236 must be fairly strong.
- d. The evidence supports the importance of precipitation (X_2), but not the index to human disturbance (X_1). This inference stems from the delta values for models 4 and 5. Note also the MLEs: the β for human disturbance is 0.14 with $se = 0.29$, whereas the β for wet/dry is 1.63 with $se = 0.30$.
- e. The β for the interaction term is nearly 0 at 0.03 with a very large $se = 0.47$; thus there is virtually no support for the interaction term.

An important issue here is that the log-likelihood value changed very little in models 6 and 7. This is the "pretending variable problem." The index to human disturbance did not improve the fit of the model (thus, $\log(\mathcal{L})$ was virtually unchanged); but the penalty for its addition was

slight (2), thus it appears to be a hypothesis/model that is "good." Indeed, it is a good model; however, this fact should not be confused with the importance of the index. Note further that model 7 with both covariates and an interaction term is also the pretentious variable problem. Here it appears that the model with both covariates and their interaction are important; however, the fact that the maximized log-likelihood changed little is the key to understanding. Again, is model 7 a good model, yes.

But this should not be taken as evidence about the importance of the interaction term. Once one understands the issues with models 6 and 7, it is seen that only model 5 has any substantial support.

Question 2. The P -value and the model probability are different entities and should not really be compared. The P -value is the probability of a test statistic as large as 6.735 or larger, given the null hypothesis is true. Only the null hypothesis is "tested" and the P -value is not a proper strength of evidence (although people misuse it as such). The alternative hypothesis is supported by default IF the null hypothesis is rejected. The alternative hypothesis is never tested.

The model probability is conditional on the data and is the probability of the model for the null hypothesis being the best model, given the data. Similarly, one also has the probability of the model for the alternative hypothesis being the best, given the data. No test statistic is involved, there is no α -level, no sampling distribution, no P -value, and no decision concerning "statistical significance." Model probabilities are a measure of the strength of evidence. Ratios of model probabilities are evidence ratios.

Question 3. Probabilities are positive quantities that sum to 1 and are usually defined as a long term frequency in repeated samples. Bayesians have extended probabilities to measure belief or uncertainty and this has been controversial. Likelihood is a relative quantity. One likelihood, by itself, is not interpretable. Likelihoods do not sum to 1 but are useful in comparing things. Likelihood and probability are very different things.

Question 4. The conditioning is twofold in this case. First, the probability is conditioned on the data. Second, it is conditional on the model set. An evidence ratio just reflects the *relative* support for any model i and any model j , regardless of what other models might be in the set.

Question 5. First, it should be remembered that R^2 is primarily meant to be a descriptive statistic; it was never meant to be used in model selection. As variables are entered into a regression model, the estimated residual variance is biased low and this makes it easier for an additional variable to enter the model (over-fitting). Using R^2 for model selection often leads to an over-fitted model.

Question 6. Consider 2 models A and B. During the derivation from K-L information to AIC there is a term (call it c) that is constant across models. Thus, we can think of the models as (trying to keep this simple and avoid notation) $c + A$ and $c + B$. When these are differenced to obtain Δ_i values, we find $(c + A) - (c + B) = c + A - c - B = A - B$. The constant is eliminated by the simple differencing. (This example assumed model B was the better of the two, assuming model A is the better model would not change the illustration.)

Question 7. She is correct in stating that the model for H_1 is estimated to be the best of the 6 models in the set. However, models for H_2 and H_3 are hardly "close," with Δ_i values of 19 and 23, respectively. These represent evidence ratios of 13,360 and 98,716, respectively. It is the difference that is important here; in this case the differences were around 20 and this is

vary large. It is often easy to think that AICc values of 3,211 and 3,234 are "similar" in magnitude and conclude (incorrectly) that they are "close."

The evidence ratio is a proper measure of the strength of evidence. Putting this into the analogy of raffle tickets is often helpful in trying to quantify the strength of evidence. In this example, the model for H_1 has 13,360 raffle tickets and the model for H_2 has one ticket. Clearly, value judgements, based on the quantitative evidence, are expected to be similar and every fair person would conclude that the hypothesis H_2 is implausible, relative to hypothesis H_1 .

Question 8.

Many practical applications focus on a simple "t-test" of a set of observations partitioned by treatment and control groups. Here, as with all experiments, the main issue is the estimation of effect size (E) -- the difference *caused* by the treatment. Using traditional methods, one often computes a "P-value," which is the probability of a test statistic as large as, or larger than, that observed, *given* the null hypothesis is true. [Many people error in thinking that a P-value is the probability that the null hypothesis is true -- this is not the proper meaning of a P-value.]

Information-theoretic approaches can be employed to provide more meaningful quantities such as,

the likelihood of both the null hypothesis and the alternative hypothesis, given the data, $\mathcal{L}(H_0|data)$ and $\mathcal{L}(H_a|data)$,

the probability of both the null hypothesis and the alternative hypothesis, given the data, $\text{Prob}(H_0|data)$ and $\text{Prob}(H_a|data)$,

the evidence ratio of the two hypotheses.

The model likelihoods and model probabilities have a rigorous, clean interpretation of the strength of evidence. [A P-value is not a measure of strength of evidence.] Model averaged estimates of effect size are also possible for observational studies.

The computation of these quantities is quite easy because a proper residual sum of squares (RSS) is available from the statistics leading to a t-statistic and the P-value. Given the RSS for each model,

$$\text{AICc} = n \cdot \log\left(\frac{\text{RSS}}{n}\right) + 2K + \frac{2K(K+1)}{n-K-1}, \quad \text{and} \quad \Delta_i = \text{AICc}_i - \text{AICc}_{\min} .$$

Model likelihoods, model probabilities and evidence ratios are easily computed from the Δ_i . The procedure for computing the RSS (using the MLEs of the structural parameters) is given below for both the classical unpaired and paired designs. Note, $\frac{\text{RSS}}{n}$ is the maximum likelihood estimate (MLE) of the residual variance.

Unpaired Design:

The null hypothesis, H_0 Effect size = 0.

$$\mu \text{ and } \sigma^2, K = 2 \text{ parameters. } \text{RSS} = \sum(x_{ci} - \hat{\mu})^2 + \sum(x_{ti} - \hat{\mu})^2$$

The alternative hypothesis, H_a Effect size = $E = \mu_c - \mu_t$.

$$\mu_c, \mu_t, \text{ and } \sigma^2, K = 3 \text{ parameters. } \text{RSS} = \sum(x_{ci} - \hat{\mu}_c)^2 + \sum(x_{ti} - \hat{\mu}_t)^2$$

Paired Design: The differences (d_i) are critical; $x_{ci} - x_{ti} = d_i$.

The null hypothesis, H_0 Effect size = 0.

$$\sigma^2, K = 1 \text{ parameters. } \text{RSS} = \sum^n (d_i)^2$$

The alternative hypothesis, H_a Effect size = $E = \bar{d}$.

$$\bar{d} \text{ and } \sigma^2, K = 2 \text{ parameters. } \text{RSS} = \sum^n (d_i - \bar{d})^2$$

These "treatment/control" data, including extensions to ANOVA designs, constitute a single data set.

Consider the example from Snedecor and Cochran (1967),

Table 4.3.1
Number of Lesions on Halves of Tobacco Leaves

Pair No.	Prepara- tion 1 X_1	Prepara- tion 2 X_2	Difference $D = X_1 - X_2$	Deviation $d = D - \bar{D}$	Squared deviation d^2
1	31	18	13	9	81
2	20	17	3	-1	1
3	18	14	4	0	0
4	17	11	6	2	4
5	9	10	-1	-5	25
6	8	7	1	-3	9
7	10	5	5	1	1
8	7	6	1	-3	9
Total	120	88	32	0	130
Mean	15	11	$\bar{D}=4$		$S_D^2=18.57$

Under the null hypothesis the RSS is 258 (from squaring and summing the 4th column in the table above; $13^2+3^2+4^2 \dots = 258$) with a sample (n) of 8 pairs. $K=1$ under the null hypothesis,

$$\begin{aligned} \text{AICc} &= n \cdot \log\left(\frac{\text{RSS}}{n}\right) + 2K + \frac{2K(K+1)}{n-K-1}, \\ &= 8 \cdot \log\left(\frac{258}{8}\right) + 2(1) + \frac{2(1)(2)}{8-1-1}, \end{aligned}$$

$$= 30.4548 .$$

Under the alternative hypothesis, $K=2$; the RSS is 130 (as shown in the table above)

$$\begin{aligned} \text{AICc} &= n \cdot \log\left(\frac{\text{RSS}}{n}\right) + 2K + \frac{2K(K+1)}{n-k-1}, \\ &= 8 \cdot \log\left(\frac{130}{8}\right) + 2(2) + \frac{2(2)(3)}{8-2-1}, \\ &= 28.7047 . \end{aligned}$$

$$\begin{aligned} \Delta_i &= \text{AICc}_i - \text{AICc}_{\min} . \\ H_o &= 30.4548 - 28.7047 = 1.7501 \\ H_a &= 28.7047 - 28.7047 = 0 . \end{aligned}$$

The best model is the alternative hypothesis H_a , estimated effect =4.

The likelihood of each model, given the data is

$$\mathcal{L}(H_o | data) = \exp(-\frac{1}{2} \Delta_i) = 0.4168$$

and

$$\mathcal{L}(H_a | data) = \exp(-\frac{1}{2} \Delta_i) = 1.0000 .$$

The model for H_a is more likely.

The probability of each model, given the data is

$$\text{Prob}(H_o | data) = 0.4168 / 1.4168 = 0.2942$$

and

$$\text{Prob}(H_a | data) = 1.0000 / 1.4168 = 0.7058 .$$

The model for H_a is more probable.

The evidence ratio is $1.0000/0.4168$ or $0.7058/0.2942 = 2.399$.

The evidence supports model H_a over model H_o . The difference is hardly overwhelming; the evidence might be judged to be weak.

These results differ from the tradition approach as Snedecor and Cochran (1967) report the t -statistic of 2.63, with 7 df, a P -value of "about 0.04" and state that the null hypothesis is rejected.

Several points can be made here. Most importantly, the P -value is not the same as the probability of model H_o ; they are not really comparable as they mean different things. The " P -value" in this case is the probability of a test statistic as large as 2.63, or larger, *given* the null hypothesis is true. P -values are a "tail probability" as they include probabilities for data more extreme than those observed. This approach gauges the probability of the data, or more extreme data, given the null is true. The P -value rests critically on the asymptotic distribution of the test statistic. The P -value should not be used as if it were a formal strength of evidence.

The model probabilities, either information-theoretic or Bayesian, provide the probability of the null model, given the data. They also provide the probability for other models; there might be only a single alternative or additional alternative models. There is no test statistic, no assumptions about the theoretical distribution of the test statistic, no concept of a cut-off (α), and no decision about "statistical significance."

In a paired or unpaired observational study one may want to model average the estimate of effect size. This is easy under an information-theoretic approach and impossible under the traditional null hypothesis testing approach. Finally, the variance of such model averaged estimates can easily incorporate a variance component due to model selection uncertainty.

ANOVA models can be similarly cast into a simple information-theoretic framework. Here again, one needs only the RSS for each model and the sample size. These quantities are always given by computer software packages. Then one computes

$$\text{AICc} = n \cdot \log\left(\frac{\text{RSS}}{n}\right) + 2K + \frac{2K(K+1)}{n-k-1},$$

$\Delta_i,$
 $\mathcal{A}(\text{model } i | \text{data}),$
 $\text{Prob}(\text{model } i | \text{data}),$ and
evidence ratios.

Regression models are easily cast into a information-theoretic framework; again one starts with the RSS.

People naturally tend to cling to traditional t -tests and ANOVAs because this is the only thing they have been taught and they are familiar with the procedure. Better methods have been developed since the early methods in the 1920s and 30s. These methods are superior in virtually all respects and their use is encouraged.

CHAPTER 5

Question 1. No. Sometimes this parameter modifies the year of the study (T) while in other models it modifies the index to human disturbance. It makes no sense to average such estimates. Once the presence of a "pretending variable" is noticed, there is little need for model averaging or incorporating a variance component for model selection uncertainty (there is little model selection uncertainty).

Question 2. The evidence ratio $E_{5,8} = 2.65$ and is a measure of the strength of evidence concerning the structure on the detection probabilities. Model 5 allows these to be modeled as a constant, $p.$, while model 8 allows these probabilities to be time-dependent, $p_t.$ The structure on the occupancy rate Ψ is the same for both models. Thus the evaluation deals just with $p.$ or $p_t.$ In this case the evidence is probably best described as weak, but there is no support for the notion that the detection probabilities are time dependent.

Question 3. The advantages is that often the first 1-3 PCs carry most of the information, thus reducing the number of regression paramaters that must be estimated. The disadvantages

frequently include the inability to interpret the result and the fact that all the original variables must still be measured.

Question 4. Burnham and Anderson (2002) explain the approach (Sect. 2.13) and offer a number of comparisons (e.g., Table 5.12).

Question 5. Perhaps there are ways to pool data from similar studies by summing either the AICc values or the Δ values? This is one of many areas needing more work.



<http://www.springer.com/978-0-387-74073-7>

Model Based Inference in the Life Sciences

A Primer on Evidence

Anderson, D.R.

2008, XXIV, 184 p. 8 illus., Softcover

ISBN: 978-0-387-74073-7