Evidence Analysis for Normal Linear Models

Brian Dennis, Professor Emeritus
Department of Fish and Wildlife Sciences
and
Department of Mathematics and Statistical Science
University of Idaho

Objective of the Evidence Project

Build model-comparison methods with improved statistical error properties to provide alternatives to Fisher-Neyman-Pearson hypothesis testing.

Approach

Combine R. Royall's concept of evidence and S. Lele's concept of evidence functions with the model selection indexes descended from H. Akaike's pathbreaking work. Extend the evidence concept to many varieties of statistical models, to models with unknown parameters, to situations with model misspecification.

Collaborators

Mark Taper, José Ponciano, Subhash Lele

José Ponciano, Mark Taper, Subhash Lele

Resources for Learning More About Evidence Statistics

In February 2020, I presented to this forum an overview of evidential statistics.

Bill Price rescued the Zoom video of that talk, and it can be downloaded here:

https://webpages.uidaho.edu/~brian/lectures/Evidence_Applied-Statistics-Seminar0220.mp4

Since then, the online journal *Frontiers of Ecology and Evolution* published a running collection of papers in the form of a special topic on evidential statistics.

The papers in the special topic feature a wide array of empirical applications, along with philosophy, theory, and new extensions. The papers have been assembled into a free pdf book which can be downloaded at the journal website or here:

https://webpages.uidaho.edu/~brian/evidence/Evidence_Statistics_2022.PDF

And now, the online journal *Entropy* has issued an invitation to submit papers to a special topic devoted to evidential statistics. Deadline for submission is Dec 31, 2023!

Evidence in Normal Linear Models

Let $\boldsymbol{y} = (y_1, y_2, \ldots y_n)'$ be an $n \times 1$ vector of observations. The normal linear model with fixed effects takes

$$\boldsymbol{y} \sim \text{MVN}(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}) \ ,$$

where $\boldsymbol{X}$ is a matrix of covariates ($n \times r$ design matrix, etc., full column rank), $\boldsymbol{\beta}$ ($r \times 1$) is a vector of unknown parameters, σ^2 is a positive parameter, and $\boldsymbol{I}$ ($n \times n$) is the identity matrix.

The likelihood function for the unknwn parameters is

$$L(\beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \text{exp}\left[-\frac{(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})}{2\sigma^2} \right] ,$$

leading to the familiar least squares ML estimates for the parameters in $\boldsymbol{\beta}$.

Hypothesis Testing

A standard formulation of statistical hypothesis testing writes

$$\boldsymbol{X\beta} = \boldsymbol{X_1\beta_1} + \boldsymbol{X_2\beta_2} ,$$

where $\boldsymbol{X_2}$ is made up of the q columns of $\boldsymbol{X}$ to be dropped from the model under the null hypothesis by setting $\boldsymbol{\beta_2} = \mathbf{0}$. We have

H_1: $\boldsymbol{\beta_2} = \mathbf{0}$ (null)
H_2: $\boldsymbol{\beta_2} \neq \mathbf{0}$ (alternative)

Writing $\widehat{\boldsymbol{\beta}}_1^*$ as the ML estimates of the $r - q$ parameters calculated under H_1, $\widehat{\boldsymbol{\beta}}$ as the ML estimates of the r parameters under H_2, the generalized likelihood ratio statistic for the test is a monotone function of an F statistic:

$$G^2 = -2\log\left(\frac{\widehat{L}_1}{\widehat{L}_2}\right) = n\log\left(1 + \frac{q}{n-r}F\right) ,$$

where

$$F = \frac{\left(\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} - \widehat{\boldsymbol{\beta}}_1^{*\prime}\boldsymbol{X}_1'\boldsymbol{y}'\right)/q}{(\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y})/(n-r)}$$ (many algebraic variants among books)

The statistic F has a noncentral F distribution: $F \sim$ F$(q, n-r, \lambda)$.

The noncentrality parameter λ is given by

$$\lambda = \frac{\boldsymbol{\beta}_2'\left(\boldsymbol{X}_2'\boldsymbol{X}_2 - \boldsymbol{X}_2'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\right)\boldsymbol{\beta}_2}{\sigma^2} .$$

Under the null model H_1, $\lambda = 0$, and F has an ordinary central F distribution.

The decision rule for the usual statistical hypothesis test would be to reject H_1 in favor of H_2 if the area to the right of the realized value of F under a central F$(q, n-r)$ distribution (i.e. the P value) is less than the desired Type I error rate α.

The power of the test would be calculated, perhaps for design purposes, with the noncentral F distribution using hypothetical true values of β_2 and σ^2 for the noncentrality parameter λ. The sample size n does not appear explicitly in the formula for λ but is implicitly contained in how the experimental units are distributed among treatments in the design matrixes $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$.

Evidence Analysis

An evidence-based approach to compare H_1 and H_2 uses an **evidence fuction** instead of a test statistic. In an evidence analysis, both models are treated the same rather than conditioning decisions on H_1. A convenient evidence function for the present linear model problem can be built from the Schwarz Information Criterion (SIC, aka BIC), one of the various information-theoretic model selection indexes.

In general, the SIC for a model i is

$$\text{SIC}_i = -2\log\left(\widehat{L}_i\right) + r_i\log(n)$$ (Schwarz 1978)

where $\widehat{L}_i$ is the maximized likelihood and r_i is the number of parameters estimated. The evidence function is the difference of the two SICs:

$$\Delta\text{SIC}_{12} = \text{SIC}_1 - \text{SIC}_2 = G^2 - \nu\log(n) \qquad (\nu = r_2 - r_1)$$

The evidence function for the normal linear model problem becomes

$\Delta\text{SIC}_{12} = n \log\left(1 + \frac{q}{n-r} F\right) - q \log(n)$.

One picks two values k_1 and k_2 for characterizing the evidence:

$\Delta\text{SIC}_{12} < k_1$ strong evidence for H_1

$k_1 < \Delta\text{SIC}_{12} < k_2$ weak or inconclusive evidence

$k_2 < \Delta\text{SIC}_{12}$ strong evidence for H_2

The values k_1 and k_2 determine two probabilities of misleading evidence M_1 and M_2:

$M_1 = \mathsf{P}(\Delta\mathsf{SIC}_{12} > k_2 \,|\, \mathsf{H}_1)$ strong but misleading evidence for H_2
$M_2 = \mathsf{P}(\Delta\mathsf{SIC}_{12} < k_1 \,|\, \mathsf{H}_2)$ strong but misleading evidence for H_1

The analysis can be pre-designed by setting the k_1 and k_2 values to attain desired low probabilities of misleading evidence. Suppose we want $M_1 \leq \gamma_1$ and $M_2 \leq \gamma_2$. $\Delta\mathsf{SIC}_{12}$ is a monotone function of the corresponding F statistic for the problem, so k_1 and k_2 can be obtained as functions of percentiles of a noncentral $\mathsf{F}(q, n-r, \lambda)$ distribution:

$$k_1 = n \log\left(1 + \tfrac{q}{n-r}\phi_1\right) - q \log(n)$$

$$k_2 = n \log\left(1 + \tfrac{q}{n-r}\phi_2\right) - q \log(n)$$

where $\mathsf{P}(F < \phi_1) = \gamma_1$ and $\mathsf{P}(F > \phi_2) = \gamma_2$.

The value of the noncentrality parameter λ to be used in the design calculations depends on the size of the departure of H_2 from H_1 that the investigator considers negligible (allowing, for instance, a decision to choose H_1 instead of H_2 in which the added complexity in H_2 is for practical purposes spurious).

The selection of a λ value can be simplified by noting that the formula for λ is in the form

$$\lambda = \frac{\boldsymbol{\beta}_2' \boldsymbol{A} \boldsymbol{\beta}_2}{\sigma^2} ,$$

where $\boldsymbol{A}$ is a $q \times q$ matrix. The quadratic form $\boldsymbol{\beta}_2' \boldsymbol{A} \boldsymbol{\beta}_2$ in the numerator is a generalized (squared) distance of $\boldsymbol{\beta}_2$ from zero. Factor out n to write

$$\boldsymbol{\beta}_2' \boldsymbol{A} \boldsymbol{\beta}_2 = n(\boldsymbol{\beta}_2' \boldsymbol{A} \boldsymbol{\beta}_2 / n) = n\delta^2$$

The quantity inside the parentheses is the generalized (squared) distance per observation. In most cases of practical interest, the sample size n inside the parentheses is absorbed into *proportions* of observations within different treatment categories. We now have λ in the following form:

$$\lambda = \frac{n\delta^2}{\sigma^2}$$

The ratio compares the per-observation generalized magnitude of β_2 to the standard deviation σ of an observation. Take δ as the largest allowable departure of β_2 from zero for use of H_1 to be acceptable and write it as a multiple of σ:

$$\delta = \nu\sigma \ .$$

For instance, if no more than half a standard deviation departure of β_2 from zero is tolerable for the use of H_1, then take $\nu = 0.5$.

The value of λ to use in the design calculations becomes

$$\lambda = n\nu^2 \ .$$

Note on the Noncentral F distribution

Users of evidence-based inference will need to get more friendly with noncentral distributions (not to mention with simulations of inferences to ascertain power-like properties). The task is made somewhat exasperating because different books and software products parameterize the distributions differently.

The noncentral F distribution denoted F(ν_1, ν_2, λ) used in these notes has a pdf given by

$$p(f) = \sum_{j=0}^{\infty} \frac{\Gamma\left(\frac{\nu_1}{2}+j+\frac{\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}+j\right)\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}+j} \left(\frac{\nu_2}{\nu_2+\nu_1 f}\right)^{\frac{\nu_1}{2}+j+\frac{\nu_2}{2}} f^{\frac{\nu_1}{2}+j-1} \frac{e^{-\frac{\lambda}{2}}\left(\frac{\lambda}{2}\right)^j}{j!} ,$$

where ν_1 and ν_2 are positive integers, and λ is a nonnegative real quantity. The mean of this distribution is

$$\mathrm{E}(F) = \frac{\nu_2(\nu_1+\lambda)}{\nu_1(\nu_2-2)} \quad \text{(provided } \nu_2 > 2) .$$

This is the noncentral F distribution in R, coded in the functions df(), pf(), rf(), and qf().

Confusion occurs because some texts (e.g. Graybill 1976) and possibly software (?) define the noncentral parameter to be $\xi = \lambda/2$. A handy way to find which distribution is contained in a software product is to simulate, say, $10,000$ observations from the distribution with $\nu_1 = 1$ and $\nu_2 = 3$, with noncentrality parameter set to a value of 2 and calculate the sample mean. If the λ parameterization is being used, the distribution mean is 9 . If the ξ parameterization is being used, the distribution mean is 15 .

Example

A 2-factor AOV concerning citrus tree fruit yield, from Ott and Longnecker (2010, example 15.8). The text and its predecessor editions were used since forever in Stat 431 (formerly Stat 401) at UI.

Data (2 observations per treatment combination):

		pesticide type 1	2	3	4
tree variety	1	49, 39	50, 55	43, 38	53, 48
	2	55, 41	67, 58	53, 42	85, 73
	3	66, 68	85, 92	69, 62	85, 99

The design matrix $\boldsymbol{X}$ for the full model has 24 rows and 12 columns, once indicator variables for interactions are included. (I used "means" coding, leave-one-out indicator variables, with intercept column included)

H_1: no interaction (six coefficients in $\boldsymbol{\beta}$ set to zero, corresponding to the interaction columns in $\boldsymbol{X}$)

H_2: interaction

NP test: $F = 1.80$, $p = .182$ fail to reject hypothesis of no interaction.

Evidence results ($n = 24$, $r = 12$, $q = 6$).

$$\Delta\text{SIC} = -3.66$$

Take $\gamma_1 = \gamma_2 = .05$. If $\gamma_1 = \gamma_2 = \gamma$, one can say that the probability of misleading evidence is no more than γ (!!)

Take $\lambda = n\nu^2$.

A. Let $\nu = 1$, so $\lambda = 24$

$k_1 = -2.16, \quad k_2 = 29.2$

Strong evidence for H_1 given that the allowable interaction strength does not exceed one standard deviation

B. Let $\nu = .5$, so $\lambda = 6$

$k_1 = -12.9, \quad k_2 = 13.3$

Inconclusive evidence for either model given that the allowable interaction strength does not exceed half a standard deviation

C. Let $\nu = .736$, so $\lambda = 13$

$k_1 = -8.48$, $\quad k_2 = 20.9$

$P(\Delta SIC < \Delta sic) \approx .18$ probability of more extreme evidence than observed (corresponds to the value of P from NP test)

Some Properties of Evidence Analysis (highlights from the *Frontiers* book)

1. As n becomes large, the probability of picking the better model approaches 1, and the error probabilities approach 0. (only one of the two error probabilities in NP testing approaches 0)

2. Evidence analysis has robustness to model misspecification built in. (NP testing can fail catastrophically with model misspecification)

3. Evidence analysis can provide support for either of the two models. (support for H_1 in NP testing is famously problematic)

4. Evidence-based interval estimates of parameter values have been derived and used.

5. The uncertainty of evidence can be assessed in the presence of model misspecification.

brian@uidaho.edu (one of the first email addresses at UI)

https://webpages.uidaho.edu/~brian/lectures/Evidence_Applied-Statistics-Seminar0220.pdf (2020 slides)

https://webpages.uidaho.edu/~brian/lectures/Evidence_Applied-Statistics-Seminar0220.mp4 (2020 video)

https://webpages.uidaho.edu/~brian/lectures/Evidence_Applied-Statistics-Seminar0123.pdf (these slides 2023)

https://webpages.uidaho.edu/~brian/lectures/Evidence_Applied-Statistics-Seminar0220.mp4 (2023 video)

https://webpages.uidaho.edu/~brian/evidence/Evidence_Statistics_2022.PDF (the book!)