Likelihood and estimation

Likelihood is the hypothetical probability that an event that has already occurred would yield a specific outcome. Likelihood differs from probability in that probability refers to future occurrences, while likelihood refers to past known outcomes.

The likelihood function is a fundamental concept in statistical inference. It indicates how likely a particular population is to produce an observed sample. The set of values that maximize the likelihood function are considered to be optimal, and are used as the estimates of the parameters.

Maximum likelihood (ML) estimation is a method of obtaining parameter estimates by optimizing the likelihood function. The likelihood function is constructed as a function of the parameters in the specified model.

Restricted maximum likelihood (REML) is an alternative methods of computing parameter estimated. REML is often preferred to ML because it produces unbiased estimates of covariance parameters by taking into account the loss of degrees of freedom that results from estimating the fixed effects in $\boldsymbol{\beta}$.

REML estimation reduces the bias in the variance component, and also handles high correlations more effectively, and is less sensitive to outliers than ML. The problem with REML for model building is that the "likelihoods" obtained for different fixed effects are not comparable. Hence it is not valid to compare models with different fixed effects using a likelihood ratio test or AIC when REML is used to estimate the model. Therefore models derived using ML must be used instead.

Assuming a statistical model $f_\theta(y)$ parameterized by a fixed and unknown set of parameters θ, the likelihood $L(\theta)$ is the probability of the observed data y considered as a function of θ (**?**).

The estimate for the fixed effects are referred to as the best linear unbiased esti-

mates (BLUE). Henderson's estimate for the random effects is known as the best linear unbiased predictor (BLUP).

0.1 Likelihood ratio test

0.1.1 Introduction

A likelihood ratio test is used to compare the fit of two models, one of which is nested within the other. This often occurs when testing whether a simplifying assumption for a model is valid, as when two or more model parameters are assumed to be related.

Likelihood ratio tests are a class of tests based on the comparison of the values of the likelihood functions of two candidate models. Likelihood ratio tests (LRTs) are a family of tests used to compare the value of likelihood functions for two models, whose respective formulations define a hypothesis to be tested (i.e. the nested and reference model).

The likelihood ratio test (LRT) is a statistical test of the goodness-of-fit between two models. A relatively more complex model is compared to a simpler model to see if it fits a particular dataset significantly better. If so, the additional parameters of the more complex model are often used in subsequent analyses.

0.1.2 Usage and Rationale

The LRT begins with a comparison of the likelihood scores of the two models:

$$LR = 2 * (lnL1 - lnL2)$$

This LRT statistic approximately follows a chi-square distribution. To determine if the difference in likelihood scores among the two models is statistically significant, we next

must consider the degrees of freedom. In the LRT, degrees of freedom is equal to the number of additional parameters in the more complex model. Using this information we can then determine the critical value of the test statistic from standard statistical tables.

For each candidate model, the '-2 log likelihood' ($M2LL$) is computed. The test statistic for each of the three hypothesis tests is the difference of the $M2LL$ for each pair of models. If the $p-$value in each of the respective tests exceed as significance level chosen by the analyst, then the null model must be rejected.

$$-2 \ln \Lambda_d = [M2LL \text{ under } H_0 \text{ model}] - [M2LL \text{ under } H_A \text{ model}] \tag{1}$$

These test statistics follow a chi-square distribution with the degrees of freedom computed as the difference of the LRT degrees of freedom.

$$\nu = [\text{ LRT df under } H_0 \text{ model}] - [\text{ LRT df under } H_A \text{ model}] \tag{2}$$

The LRT is explained in more detail by Felsenstein (1981), Huelsenbeck and Crandall (1997), Huelsenbeck and Rannala (1997), and Swofford et al. (1996).

Both models are fitted to the data and their log-likelihood recorded. The test statistic (usually denoted D) is twice the difference in these log-likelihoods:

The model with more parameters will always fit at least as well (have a greater log-likelihood). Whether it fits significantly better and should thus be preferred can be determined by deriving the probability or p-value of the obtained difference D.

The test requires nested models, that is, models in which the more complex one can be transformed into the simpler model by imposing a set of linear constraints on the parameters.

In a concrete case, if model 1 has 1 free parameter and a log-likelihood of 8012 and the alternative model has 3 degrees of freedom and a LL of 8024, then the probability of this difference is that of chi-square of 24 = 2(8024 8012) under 2 = 3 1 degrees of freedom. Certain assumptions must be met for the statistic to follow a chi-squared distribution and often empirical p-values are computed.

0.1.3 LRT Test Statistic

The test statistic for each of the three hypothesis tests is the difference of the $M2LL$ for each pair of models. The test statistic for the LRT is the difference of the log-likelihood functions, multiplied by -2.

The probability distribution of the test statistic is approximated by the χ^2 distribution with $(\nu_1 - \nu_2)$ degrees of freedom, where ν_1 and ν_2 are the degrees of freedom of models 1 and 2 respectively.

If the $p-$value in each of the respective tests exceed as significance level chosen by the analyst, then the null model must be rejected.

The significance of the likelihood ratio test can be found by comparing the likelihood ratio to the χ^2 distribution, with the appropriate degrees of freedom.

$$\nu = [\text{ LRT df under } H_0 \text{ model}] - [\text{ LRT df under } H_A \text{ model}] \tag{3}$$

L= - 2ln is approximately distributed as 2 under H_0 for large sample size and under the normality assumption.

$$-2 \ln \Lambda_d = [M2LL \text{ under } H_0 \text{ model}] - [M2LL \text{ under } H_A \text{ model}] \tag{4}$$

These test statistics follow a chi-square distribution with the degrees of freedom computed as the difference of the LRT degrees of freedom.

In many cases, the probability distribution of the test statistic can be approximated by a chi-square distribution with (df1 df2) degrees of freedom, where df1 and df2 are the degrees of freedom of models 1 and 2 respectively.

0.1.4 Statistical Assumptions for Likelihood Ratio Tests

We generally use LRTs to evaluate the significance of terms in the random effects structure, i.e. different nested models are fitted in which the random effects structure is changed.

The LRT is only valid if used to compare hierarchically nested models. That is, the more complex model must differ from the simple model only by the addition of one or more parameters. Adding additional parameters will always result in a higher likelihood score. However, there comes a point when adding additional parameters is no longer justified in terms of significant improvement in fit of a model to a particular dataset. The LRT provides one objective criterion for selecting among possible models.

When testing hypotheses around covariance parameters in an LME model, REML estimation for both models is recommended by (West et al., 2007), as it REML estimation can be shown to reduce the bias inherent in ML estimates of covariance parameters. Conversely, Pinheiro and Bates (1994) advises that testing hypotheses on fixed-effect parameters should be based on ML estimation, and that using REML would not be appropriate in this context.

A general method for comparing nested models fit by maximum likelihood is the ***likelihood ratio test***. This test can be used for models fit by REML (restricted maximum liklihood), but only if the fixed terms in the two models are invariant, and

both models have been fit by REML. Otherwise, the argument: method="ML" must be employed (ML = maximum likelihood).

Generally, likelihood ratio tests should be used to evaluate the significance of terms on the random effects portion of two nested models, and should not be used to determine the significance of the fixed effects. A simple way to more reliably test for the significance of fixed effects in an LME model is to use conditional F-tests, which will give the most reliable test of the fixed effects included inthe model.

0.1.5 Testing Procedures

Roy's methodology requires the construction of four candidate models. The first candidate model is compared to each of the three other models successively. It is the alternative model in each of the three tests, with the other three models acting as the respective null models.

Likelihood ratio tests are very simple to implement in `R`, simply use the '`anova()`' commands. Sample output will be given for each variability test.

A general method for comparing nested models fit by maximum liklihood is the liklihood ratio test. This test can be used for models fit by REML (restricted maximum liklihood), but only if the fixed terms in the two models are invariant, and both models have been fit by REML. Otherwise, the argument: method=ML must be employed (ML = maximum liklihood).

Example of a liklihood ratio test used to compare two models:

CODE HERE

Generally, liklihood ratio tests should be used to evaluate the significance of terms on the random effects portion of two nested models, and should not be used to determine the significance of the fixed effects.

0.1.6 Relevance of Estimation Methods

The problem with REML for model building is that the "likelihoods" obtained for different fixed effects are not comparable. Hence it is not valid to compare models with different fixed effects using a likelihood ratio test or AIC when REML is used to estimate the model. Therefore models derived using ML must be used instead.

Nested LME models, fitted by ML estimation, can be compared using the likelihood ratio test (**?**). Models fitted using REML estimation can also be compared, but only if both were fitted using REML, and both have the same fixed effects specifications.

Likelihood ratio tests are generally used to test the significance of terms in the random effects structure.

For both REML and ML estimates, the nominal $p-$values for the LRT statistics under a χ^2 distribution with 2 degrees of freedom are much greater than empirical values. A number of ways of dealing with this issues are discussed (**?**, pg.86).

One should be aware that these p-values may be conservative. That is, the reported p-value may be greater than the true p-value for the test and, in some cases, it may be much greater.(**?**, pg.87).

Pinheiro & Bates (2000; p. 88) argue that Likelihood Ratio Test comparisons of models varying in fixed effects tend to be anticonservative i.e. will see you observe significant differences in model fit more often than you should.

0.1.7 Score Function and Fisher Information

Such a test can also be used for models fitted using REML, but only if both models have been fitted by REML, and if the fixed effects specification is the same for both models.

Each of these three test shall be examined in more detail shortly. The power of the

likelihood ratio test may depends on specific sample size and the specific number of replications, and Roy (2009) proposes simulation studies to examine this further.

The score function $S(\theta)$ is the derivative of the log likelihood with respect to θ,

$$S(\theta) = \frac{\partial}{\partial\theta} l(\theta),$$

and the maximum likelihood estimate is the solution to the score equation $S(\theta) = 0$. The Fisher information $I(\theta)$, which is defined as

$$I(\theta) = -\frac{\partial^2}{\partial\theta^2} l(\theta),$$

give rise to the observed Fisher information ($I(\hat{\theta})$) and the expected Fisher information ($\mathcal{I}(\theta)$).

0.2 Nested Models

0.2.1 Definitions of Nested Models

An important step in the process of model selection is to determine, for a given pair of models, if there is a "nesting relationship" between the two.

We define Model A to be "nested" in Model B if Model A is a special case of Model B, i.e. Model B with a specific constraint applied.

One model is said to be *nested* within another model, i.e. the reference model, if it represents a special case of the reference model (Pinheiro and Bates, 1994).

0.2.2 Nesting: Model Selection Using Likelihood Ratio Tests

The relationship between the respective models presented by Roy (2009) is known as "nesting". Hypotheses can be formulated in the context of a pair of models that have

a nesting relationship West et al. (2007). An important step in the process of model selection is to determine, for a given pair of models, if there is a "nesting relationship" between the two.

Model A to be nested in the reference model, *Model B*, if *Model A* is a special case of *Model B*, or with some specific constraint applied. One model is said to be *nested* within another model, i.e. the reference model, if it represents a special case of the reference model (Pinheiro and Bates, 1994).

LRTs can be used to test hypotheses about covariance parameters or fixed effects parameters in the context of LMEs.

0.2.3 Nested and Reference Models

Hypotheses can be formulated in the context of a pair of models that have a nesting relationship (West et al., 2007).

LRTs are a class of tests used to compare the value of likelihood functions for two models defining a hypothesis to be tested (i.e. the nested and reference model).

The significance of the likelihood ratio test can be found by comparing it to the χ^2 distribution, with the appropriate degrees of freedom.

0.3 Other material

0.3.1 Implementing Likelihood Ratio Tests using `R`

- Example of a likelihood ratio test used to compare two models:

  ```
  >anova(modelA, modelB)
  ```

- The output will contain a p-value, and this should be used in conjunction with the AIC scores to judge which model is preferred. Lower AIC scores are better.

- Generally, likelihood ratio tests should be used to evaluate the significance of terms on the random effects portion of two nested models, and should not be used to determine the significance of the fixed effects.

- A simple way to more reliably test for the significance of fixed effects in an LME model is to use conditional F-tests, as implemented with the simple "anova" function.

  Example:

  `>anova(modelA)`

  will give the most reliable test of the fixed effects included in model1.

A simple way to more reliably test for the significance of fixed effects in an LME model is to use conditional F-tests, as implemented with the simple anova function.

```
> anova(MCS1,MCS2)


Model df    AIC    BIC  logLik   Test L.Ratio p-value
MCS1     1  8 4077.5 4111.3 -2030.7
MCS2     2  7 4075.6 4105.3 -2030.8 1 vs 2 0.15291  0.6958
```

The output will contain a p-value, and this should be used in conjunction with the AIC scores to judge which model is preferred. Lower AIC scores are better.

0.3.2 Pinheiro Bates

A general method for comparing nested models fitted by ML is the ***likelihood ratio test*** (Lehmann, 1986). Such a test can also be used for models fitted using REML, but

only if both models have been fitted by REML, and if the fixed effects specification is the same for both models.

If k_i is the number of parameters to be estimated in model i, then the asymptotic, or "large sample", distribution of the LRT statistic, under the null hypothesis that the restricted model is adequate, is a χ^2 distribution with $k_2 - k_1$ degrees of freedom (Pinheiro and Bates, 1994, pg.83).

We generally use LRTs to evaluate the significance of terms in the random effects structure, i.e. different nested models are fitted in which the random effects structure is changed.

0.3.3 Akaike Information Critierion

Akaike (1974) introduces the Akaike information criterion (AIC), a model selection tool based on the likelihood function. The AIC is a model selection method, assessing how the goodness of fit of a model. It is computed as follows:

$$AIC = -2l_{max} + 2k$$

with l_{max} as the log-likelihood maximum and k as the number of parameters. Given a data set, candidate models are ranked according to their AIC values, with the model having the lowest AIC being considered the best fit.

Additionally nested models may be compared by using the Akaike Information Criterion,(AIC) and the Bayesian Information Criterion (BIC).

0.3.4 LRTs for covariance parameters

West et al. (2007) recommends that, when testing hypotheses around covariance parameters in an LME model, REML estimation for both models should be used. REML

estimation can be shown to reduce the bias inherent in ML estimates of covariance parameters (Morrell, 1998).

LRTs can be used to test hypotheses about covariance parameters or fixed effects parameters in the context of LMEs. Each of these three test shall be examined in more detail shortly.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723.

Lehmann, E. (1986). Testing statistical hypotheses wadsworth & brooks. *Cole, Pacific Grove, California*.

Morrell, C. H. (1998). Likelihood ratio testing of variance components in the linear mixed-effects model using restricted maximum likelihood. *Biometrics*, 1560–1568.

Pinheiro, J. and D. Bates (1994). *Mixed Effects Models in S and S plus* (2nd ed.). Reading, Massachusetts: Springer.

Roy, A. (2009). An application of the linear mixed effects model to ass the agreement between two methods with replicated observations. *Journal of Biopharmaceutical Statistics 19*, 150–173.

West, B., K. Welch, and A. Galecki (2007). *Linear Mixed Models: a Practical Guide Using Statistical Software*. Chapman and Hall CRC.