Document information

- *Title:* An alternative approach to psedu-likelihood model selection in the generalized linear mixed modeling framework
- *Authors:* Patrick Ten Eyck, Joseph E. Cavanaugh
- *DOI:* `https://doi.org/10.1007/s13571-017-0130-5`
- *File name:* `Ten_Eyck_Cavanaugh_2018a.pdf`

Introduction

In the maximum likelihood framework, the use of information criteria based on the empirical likelihood is pervasive, where the empirical likelihood appears in the goodness-of-fit (GOF) term of the criterion. To compare such criteria across different fitted models, the outcome data must be identical; otherwise, the likelihoods will not correspond.

The pseudo-likelihood method is a conventional fitting approach for the framework of the GLMM. With this method, pseudo-data are generated via a transformation of the outcome and used as a surrogate for the original response data.

- Pseduo-data are derived from a Taylor series expansion that utilizes both constructs from the candidate model and the original outcome. The purpose of this expansion is to offer a new outcome with an approximate normal distribution.
- Pseudo-data are inconsistent for different model specifications, leading to pseudo-likelihoods that are not comparable; thus, selection criteria based on the resulting GOF statistics are fundamentally dissimilar (i.e., no comparisons possible).

Big idea of new method: To ensure that use of model selection criteria is valid, the same pseudo-data should be used for all models under consideration. The simplest way to accomplish this objective is to use the full model to obtain the pseduo-data, and to subsequently fit all subset candidate models with this generated outcome.

Background

Model selection criteria

An optimal statistical model is characterized by three features:

1. parsimony, which refers to model simplicity;
2. goodness-of-fit (GOF), which indicates the conformity of the fitted model to the data at hand;
3. generalizability, which reflects the ability of the fitted model to predict or describe new outcomes.

Parsimony and GOF tend to pull in opposing directions with regards to model complexity, so it is important to strike a suitable balance between those two attributes, while still achieving generalizability.

There are two types of model selection criteria.

1. Assuming that the generating model is of an infinite dimension and thus is not in the candidate collection, an *efficient* criterion will asymptotically select the fitted candidate model which minimizes the mean squared error of prediction.
    - Criterion example: AIC

2. Suppose that the generating model is of a finite dimension, and that this model is represented in the candidate collection under consideration. A *consistent* criterion will asymptotically select the fitted candidate model having the correct structure with probability one.
    - Criterion example: BIC

An advantage of model selection criteria over common model selection comparison inferential techniques, such as the likelihood ratio test, is that the models under consideration do not need to be nested or even follow the same distribution. As long as the models are fitted using the same outcome, the selection criteria can be compared to determine the more appropriate fit.

Problem with pseudo-likelihood criteria

See the notes taken in the **Introduction**.

Generalized linear mixed models

GLMMs are the best available recourse for analyzing normal and non-normal data that involve random effects, requiring only the specification of a conditional distribution for the outcome, a link function, and a covariance structure for the random effects.

Pseudo-likelihood fitting approach

Primary fitting approaches for GLMMs are pseudo-likelihood, Gaussian quadrature, and Laplace approximation method.

- The advantage of using pseudo-likelihood over these alternate approaches is its computational efficiency.
- "pseduo" is used because the likelihood is based on a linearized transformatioin of the data, which is assumed normal and is not the actuall likelihood based on the original data and its underlying distribution.

The default approach for the `GLIMMIX` procdure fits GLMMs based on linearization, which utilizes Taylor series expansions to approximate the data using pseudo-data.

- Through the iterative process, the pseudo-data are constructed using current regression and covariance parameter estimates.
- The GLMM is then approximaed by a LMM based on the pseudo-data.
- The empirical pseudo-likelihood under the GLMM framework can seemingly be used to calculate model selection criteria in a similar manner to the empirical likelihood under the GLM framework.
- Since model selection criteria constructed under the pseudo-likelihood approach utilize a different pseudo-data vector for each model under consideration, comparisons of conformity between the models and the data are inappropriate.

Investigative simulation

Results from the simulation are shown in Table 1, where all selection criteria based on the pseduo-likelihood selects the wrong model from the candidate class of models.

Proposed solution

See **Big idea of new method** in the **Introduction**

New method

Heuristic justification

Ideally, we would construct the pseudo-data based on the true model, which we do not know. However, if we assume that we have access to the predictors in the true model, we can generate the pseudo-data by using the full model, which subsumes the true model. Using this full model pseudo-data, a LMM can be fit with an subset of predictors from the full model.

- Since all models will share the same pseudo-data, information criteria can validly be compared for the purposes of model selection.

Implementation via SAS `PROC MIXED/PROC GLIMMIX`

- Fit the full model using `GLIMMIX` and output the predicted and residual components of the pseudo-data.
- Use the `MIXED` procedure with the full model pseudo-data to fit all candidate models of interest and generate the desired information criteria.

Simulation study

The proposed method tends to do better when the outcome is either binomial, poisson, or gamma. For a bernoulli outcomes, it is not bad, but it is not great when compared to the true model. This makes sense because the normal distribution poorly describes the pseudo-data based on a Bernoulli resposne.