Test (2009) 18: 248–255
DOI 10.1007/s11749-009-0151-0

DISCUSSION

Comments on: Goodness-of-fit tests in mixed models

Jiming Jiang · Thuan Nguyen

Published online: 14 May 2009

Professors Claeskens and Hart's paper provides a comprehensive and very interesting review on goodness-of-fit tests for mixed model diagnostics. The review has generated some renewed interest on a problem that one of us tried to address eight years ago. Our discussion focuses on four specific topics, namely, the importance of generalized linear mixed model (GLMM) diagnostics, the model selection approach to goodness-of-fit tests, the simulation example, and bin selection in χ^2 goodness-of-fit tests.

1 GLMM diagnostics

The review mainly focuses on linear mixed model diagnostics. It may be argued, however, that model diagnostics is even more important for GLMMs than for linear mixed models. For example, it is well known that the Gaussian maximum likelihood (ML) and restricted maximum likelihood (REML) estimators of the fixed-effects and variance components in a linear mixed model remain consistent even if the normality assumption is violated (Jiang 1996). On the other hand, the distributional assumption on the random effects is critically important for the consistency of MLEs in GLMM. To see this, consider the following simple example.

This comment refers to the invited paper available at doi:10.1007/s11749-009-0148-8.

J. Jiang (✉) · T. Nguyen
University of California, Davis, CA, USA
e-mail: jiang@wald.ucdavis.edu

J. Jiang · T. Nguyen
Davis and Oregon Health and Science University, Partland, OR, USA

Example 1 Suppose that, given the random effects $\xi_1, \ldots, \xi_n$ which are i.i.d. with mean 0 and variance 1, the responses $y_1, \ldots, y_n$ are conditionally independent such that $y_i \sim \text{Bernoulli}(p_i)$, where $\text{logit}(p_i) = \mu + \xi_i$, and μ is an unknown parameter. First note that the Gaussian MLE under the corresponding linear mixed model, $y_i = \mu + \xi_i$, $i = 1, \ldots, n$, is the sample mean, $\hat{\mu} = \bar{y} = n^{-1} \sum_{i=1}^{n} y_i$, which is consistent regardless of the actual distribution of ξ. But let us see what happens under the current logistic model. Under the assumption that the ξ_i's are normal, the likelihood function can be expressed as

$$L(\mu) = \prod_{i=1}^{n} \mathrm{E}\big[\exp\big\{y_i(\mu + \xi) - \log\big(1 + e^{\mu+\xi}\big)\big\}\big], \tag{1}$$

where the expectations are taken with respect to $\xi \sim N(0, 1)$. The likelihood equation, obtained by taking the logarithm of (1) and setting the derivative equal to zero, is equivalent to the following:

$$N_1 \mathrm{E}\left(\frac{1}{1 + e^{\mu+\xi}}\right) - N_0 \mathrm{E}\left(\frac{e^{\mu+\xi}}{1 + e^{\mu+\xi}}\right) = 0, \tag{2}$$

where $N_j = \#\{1 \le i \le n : y_i = j\}$, $j = 0, 1$. Now suppose that the true distribution of ξ_i is, actually, a two-point distribution: $\mathrm{P}(\xi_i = -1) = \mathrm{P}(\xi_i = 1) = 1/2$. It follows that the marginal distribution of y_i is given by

$$\mathrm{P}(y_i = j) = \begin{cases} \frac{1}{2}\left(\frac{1}{1+e^{\mu-1}} + \frac{1}{1+e^{\mu+1}}\right), & j = 0, \\ \frac{1}{2}\left(\frac{e^{\mu-1}}{1+e^{\mu-1}} + \frac{e^{\mu+1}}{1+e^{\mu+1}}\right), & j = 1. \end{cases} \tag{3}$$

With this, it can be shown that the likelihood equation (2) is biased in the sense that the expectation of the left side is not equal to zero, in general, which leads to inconsistent estimator of μ (e.g., White 1982). To see this, simply take the expectation of the left side with respect to the true distribution of the y_i's given by (3). This leads to the equation

$$\mathrm{E}\left(\frac{e^{\mu+\xi}}{1 + e^{\mu+\xi}}\right) = \frac{1}{2}\left(\frac{e^{\mu-1}}{1 + e^{\mu-1}} + \frac{e^{\mu+1}}{1 + e^{\mu+1}}\right). \tag{4}$$

Let $\psi(\mu)$ denote the difference of the two sides of (4). Using Monte Carlo integration, a plot of $\psi(x)$ against x for positive x's is shown in Fig. 1. It is clear that, with the exception of a few values (actually $\mu = 0$ is a root to the equation, and there is another root between 2 and 3), $\psi(\mu)$ is not identical to zero.

Unfortunately, the literature on GLMM diagnostics is even (much) more sparse than that for linear mixed model diagnostics. Recently, Gu (2008) extended the χ^2 goodness-of-fit test of Jiang et al. (2001) to GLMM diagnostics. It may be argued that the χ^2 test (see later discussion) is more suitable for GLMM diagnostics than for linear mixed model diagnostics. One reason is that the selection of bins for the χ^2 test is a difficult problem. On the other hand, in some cases of GLMMs, there are natural choices of the bins. For example, if the responses are binomial with possible

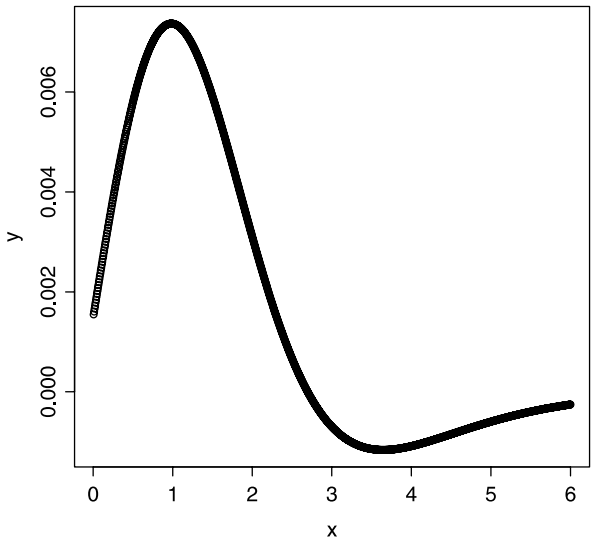

Fig. 1 Plot of $y = \psi(x)$ against x

values $1, \ldots, k$, then the natural choice of bins consist exactly with these values; in the case of count data, the natural bins are $0, 1, 2, \ldots, k$ and any values larger than k, for some k.

2 The model selection approach

The model selection approach for goodness-of-fit tests discussed by Professors Claeskens and Hart is, indeed, very interesting. The idea is to expand the underlying density function of the random effects around the standard normal density. The expansion is in the form of the normal density multiplied by a polynomial. The degree of the polynomial is then selected by the information criteria, such as BIC.

It appears that the method is only suitable for testing the null hypothesis that the distribution of the random effects is normal. Of course, this is the case in most applications, but there are exceptions, for example, in the case of frailty models (e.g., Fu et al. 2002).

The consistency of BIC and HQ is mentioned; however, such a concept applies only to the situation where the true underlying model is of finite dimension and among the candidate models. In the current situation, however, the true underlying model may not be any of the approximating models (i.e., normal density multiplied by a polynomial; see above). In such a case, it may be argued that the BIC and HQ are inconsistent, while the AIC is consistent.

Furthermore, the information criteria apply to the so-called conventional situations where the effective sample size can be easily determined. However, as noted by Jiang et al. (2008), such conventional situations do not include the case of mixed models. We use the following example given by the latter authors to illustrate the problem.

Example 2 Consider a model with crossed random effects,

$$y_{ij} = \mu + u_i + v_j + e_{ij},$$

$i = 1, \ldots, m_1$, $j = 1, \ldots, m_2$, where u_is, v_js are random effects, and e_{ij}s are errors. It is assumed that u_is are i.i.d. with mean 0 and variance σ_u^2; v_js are i.i.d. with mean 0 and variance σ_v^2; e_{ij}s are i.i.d. with mean 0 and variance σ_e^2, and u, v, e are independent. It is well known that, in this case, the effective sample size for estimating σ_u^2 is m_1 (not $n = m_1 m_2$, the total sample size); similarly, the effective sample size for estimating σ_v^2 is m_2. Now suppose that one attempts to use the BIC to select the order of the polynomial for testing a hypothesis regarding the distributions of both random effects, u and v. It is not clear what is the effective sample size (m_1, m_2, $m_1 + m_2$, or $m_1 m_2$?) that is needed for BIC (in order to compute the penalty term $\log N$, where N is supposed to be the effective sample size).

On the other hand, the fence methods proposed by Jiang et al. (2008) is suitable for mixed model selection and other nonconventional model selection problems. It would be interesting to see how the methods apply to the goodness-of-fit tests for mixed models.

3 The simulation example

The case of the simulated example in Sect. 6.3 is a bit unusual in that the range of Y_{ij} is dominated by that of x_{ij} and there is little variation due to the random effects and errors. For example, the range of Y_{ij} is taken as $[0, 22]$, of which $[0, 21]$ is due to the range of $\beta_0 + \beta_1 x_{ij}$, and only plus/minus 1 around 21, that is, the interval $[20, 22]$ is likely to be influenced by $\gamma_i + \epsilon_{ij}$. Intuitively, Pearson's test, which depends heavily on the range of the observations, is not expected to do well in such a case.

A potential remedy for this kind of observations [note that, in practice, the small influence of random effects and errors in a linear mixed model are indicated by the estimated variance components, whose consistency does not require the normality assumption (Jiang 1996)] is to standardize, or studentize, the observations before applying Pearson's test. Under the null hypothesis, we have

$$Z_{ij} = \frac{Y_{ij} - \beta_0 - \beta_1 x_{ij}}{\sqrt{\sigma_\gamma^2 + \sigma_\epsilon^2}} \sim N(0, 1). \tag{5}$$

Therefore, we consider

$$\hat{Z}_{ij} = \frac{Y_{ij} - \hat{\beta}_0 - \hat{\beta}_1 x_{ij}}{\sqrt{\hat{\sigma}_\gamma^2 + \hat{\sigma}_\epsilon^2}}, \tag{6}$$

where $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_\gamma^2, \hat{\sigma}_\epsilon^2$ are the REML estimators of the corresponding parameters, and call (6) the studentized observation. We then define $\hat{O}_k = \sum_{i,j} 1_{(\hat{Z}_{ij} \in I_k)}$, where I_k is the kth interval, or bin, $1 \le k \le M$, and M is the total number of bins. The expected frequencies are computed under (5) [which are asymptotically correct under (6) and

the null hypothesis], that is, $E_k = nm\{\Phi(a_k) - \Phi(a_{k-1})\}$, where $a_{k-1} < a_k$ are the end points of I_k (open or closed), and $\Phi(\cdot)$ is the cdf of $N(0, 1)$. Note that the E_k's involve no estimated parameters. The new test statistic is then

$$\tilde{\chi}^2_M = \frac{1}{nm^2}\sum_{k=1}^{M}(\hat{O}_k - E_k)^2. \tag{7}$$

In contrast, the original test statistic is

$$\hat{\chi}^2_M = \frac{1}{nm^2}\sum_{k=1}^{M}(O_k - \hat{E}_k)^2. \tag{8}$$

It is clear that the difference is where to put the hat. For (8), the O_k's are observed while the E_k's estimated. For (7), it is the other way around: the O_k's are estimated, while the E_k's are known. It should be pointed out that the two test statistics may not have the same asymptotic null distribution (e.g., Jiang et al. 2001).

Also, in practice it is preferable to compute the critical values via a Monte Carlo, or bootstrap method. The reason is that the critical value depends on the computation of an asymptotic covariance matrix (Jiang et al. 2001), whose analytic form is complicated even for a fairly simple case. Monte Carlo or bootstrap method is much easier to operate and, more importantly, can avoid analytic and coding errors in computing the critical value. To compute the bootstrapped critical value, simply simulate the data under the null hypothesis, with the unknown parameters estimated by, say, their (Gaussian) REML estimators. Once again, the result of Jiang (1996) guarantees that these estimators are consistent even without the null hypothesis. Then, we compute the test statistic [(7) or (8)] under the repeated bootstrap samples and obtain the bootstrapped critical value at a given significance level.

We carry out a small simulation study under the same setting of the simulated example. It should be pointed out that it is not very clear how the random effects are generated in Professors Claeskens and Hart's simulations. A condition for Pearson's test is that the random effects have finite fourth moments (Jiang et al. 2001). However, the Cauchy distribution (t_1) does not even have a mean, so it is not clear how the random effect γ is generated so that it has mean zero and variance $\sigma^2_\gamma = 0.1$. So we have to leave this one aside. The normal mixture distribution is also not very clear, but somehow we can deal with it, assuming that this is also what is done in the paper. The normal mixture of $N(-4, 0.1)$ with probability 0.1 and $N(4, 0.1)$ with probability 0.9 has mean 3.2 (not 0; note that the random effects need to have mean zero in order for β_0 to be identifiable) and variance 5.86 (not 0.1, as indicated). What we do is to standardize the distribution by $\gamma = (X - 3.2)/\sqrt{58.6}$, where X has the above normal mixture distribution, so that γ has mean 0 and variance 0.1. We compare the performance of the original test (8), Test I, with two different versions of (7). The first, Test II, uses the equal-length method to determine the bins, given M (the range of $N(0, 1)$ is taken as $[-4, 4]$); the second, Test III, uses the equal-probability method to determine the bins, given M (so that $E_k = 1/M$ for all k). The results based on 100 simulation runs are reported in Table 1.

Table 1 Simulation comparisons of χ^2 tests

# of bins	Test I		Test II		Test III	
	$\alpha = .10$	$\alpha = .05$	$\alpha = .10$	$\alpha = .05$	$\alpha = .10$	$\alpha = .05$
3	.16	.08	.25	.12	.28	.20
4	.11	.05	.42	.35	.27	.17
5	.10	.06	.30	.17	.26	.20
6	.11	.05	.40	.28	.25	.17
7	.08	.03	.39	.25	.65	.62
8	.06	.05	.29	.21	.30	.19
9	.09	.02	.34	.24	.24	.16
10	.15	.07	.31	.20	.26	.19

It is seen that for $M = 3$, which is the case considered by Professors Claeskens and Hart, the empirical (or observed) powers for the non-studentized test [i.e., (8)] are quite low (but not as low as zero, as indicated by Table 2 of Claeskens and Hart's paper). The studentized test [i.e., (7)] works better, with either equal-length or equal-probability bins (the latter seems to perform better for $M = 3$). This confirms our earlier speculation that the poor performance of Pearson's test may be partially due to the fact that the range of Y_{ij} is dominated by the fixed effects in this example. Another factor that definitely plays a role is the number of bins, M. It is seen that, with a suitable choice of M, the empirical power can be as high as .42 for Test II and .65 for Test III. The question is how to choose the "suitable" M. We continue our discussion in the next section.

4 Bin selection in χ^2 tests

First we would like to present a plot that shows the differences in the empirical power of the studentized test (both Test II and Test III) when M changes. See Fig. 2. In each case the plots show a similar pattern. This, once again, raises an old (and difficult) question: How to select the number of bins for Pearson's χ^2 test?

The guidelines given in Jiang et al. (2001) in choosing M, that is, $M = [n^{1/5}]$, is based on an asymptotic result of Sanatov (1980). However, this needs not give an optimal choice in a finite sample situation. On the other hand, the choice of M may be viewed as a model selection problem. It is not clear, however, how the information criteria can play a role in this regard. Note that this is not model selection in the traditional sense, especially when the observations are correlated. However, the problem appears to fit naturally with the scope of the fence methods (Jiang et al. 2008). The latter finds a natural way of balancing "model fit" and "model complexity." Note that in a testing problem one also needs to balance two things: size and power. If one defines the measure of "lack-of-fit," Q_M, in the fence method as the probability of rejection under the null hypothesis given M, where M represents the number of, say, equal-probability and/or equal-length bins, then all the χ^2 tests that are designed at a

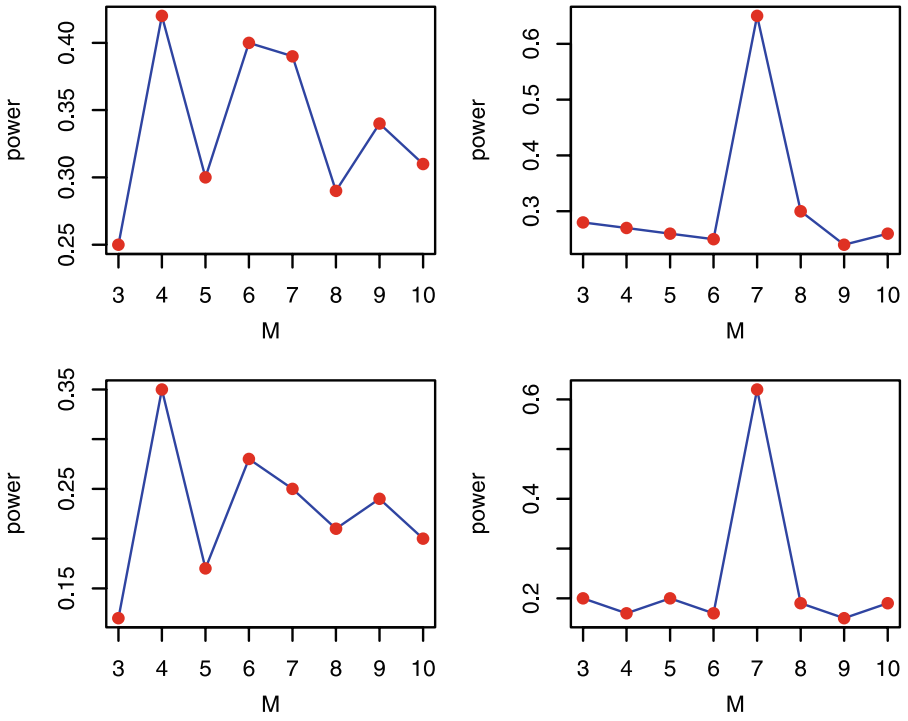

Fig. 2 Empirical power against number of bins. *Left plots*: Test II ($\alpha = .10$ at *top* and .05 at *bottom*). *Right plots*: Test III ($\alpha = .10$ at *top* and .05 at *bottom*)

given significance level, α, are "in the fence" in the sense that they satisfy the fence inequality

$$Q_M - Q_1 \leq c \tag{9}$$

with $c = \alpha$ if Q_1 is understood as 0 (note that when $M = 1$, the χ^2 statistic is identical to zero and therefore cannot reject any hypothesis). Now it is up to us to define a criterion of optimality to select a model within the fence (this is another feature of the fence, that is, the criterion of optimality for selecting a model within the fence is flexible). Naturally, the criterion is optimal power. Of course, this is something that one would have come up with anyway without using the fence. The question is how to do this.

Note that the power is always calculated under a given alternative. For testing for normality, the alternative may be chosen as one of the approximating distributions considered by Professors Claeskens and Hart, which is the standard normal density multiplied by a polynomial of a certain degree. The degree of the polynomial may be chosen by the information criterion, as suggested by Professors Claeskens and Hart or, again, by the fence method (Jiang et al. 2008). Once the alternative is determined, one can estimate the parameters under the alternative distribution using, say, the maximum likelihood, and then use a model-based (or parametric) bootstrap method to draw samples under the alternative distribution, and hence evaluate the

empirical power of the χ^2 test for each given M. The optimal M corresponds to the one with the maximum empirical power.

References

Fu P, Rao JS, Jiang J (2002) Robust estimation of multivariate failure data with time-modulated frailty. J Mod Appl Stat Methods 1:367–378

Gu Z (2008) On the diagnostics of generalized linear mixed models. PhD dissertation, Dept Statist, Univ of Calif, Davis, CA

Jiang J (1996) REML estimation: asymptotic behavior and related topics. Ann Stat 24:255–286

Jiang J, Lahiri P, Wu C (2001) A generalization of the Pearson's χ^2 goodness-of-fit test with estimated cell frequencies. Sankhya 63A:260–276

Jiang J, Rao JS, Gu Z, Nguyen T (2008) Fence methods for mixed model selection. Ann Stat 36:1669–1692

Sanatov VV (1980) Uniform estimates of the rate of convergence in the multi-dimensional central limit theorem. Theory Probab Appl 25:745–759

White H (1982) Maximum likelihood estimation of misspecified models. Econometrica 50:1–25