A simulation-based goodness-of-fit test for random effects in generalized linear mixed models

Waagepetersen, Rasmus Plenge

Publication date:
2005

Document Version
Publisher's PDF, also known as Version of record

AALBORG UNIVERSITY

A simulation-based goodness-of-fit test for random effects in generalized linear mixed models

by

Rasmus Waagepetersen

R-2005-19 Maj 2005

Department of Mathematical Sciences
Aalborg University
Fredrik Bajers Vej 7 G ▪ DK - 9220 Aalborg Øst ▪ Denmark
Phone: +45 96 35 80 80 ▪ Telefax: +45 98 15 81 29
URL: http://www.math.aau.dk

ISSN 1399–2503 ▪ On-line version ISSN 1601–7811

A simulation-based goodness-of-fit test for random effects in generalized linear mixed models

Rasmus Waagepetersen
Institute of Mathematical Sciences, Aalborg University
Fredrik Bajersvej 7G, DK-9220 Aalborg

Abstract

The goodness-of-fit of the distribution of random effects in a generalized linear mixed model is assessed using a conditional simulation of the random effects conditional on the observations. Provided that the specified joint model for random effects and observations is correct, the marginal distribution of the simulated random effects coincides with the assumed random effects distribution. In practice the specified model depends on some unknown parameter which is replaced by an estimate. We obtain a correction for this by deriving the asymptotic distribution of the empirical distribution function obtained from the conditional sample of the random effects. The approach is illustrated by simulation studies and data examples.

Keywords: conditional simulation, empirical distribution function, generalized linear mixed model, goodness-of-fit, random effects.

Running header: A simulation-based goodness-of-fit test.

1 Introduction

This paper is concerned with assessment of the distributional assumptions for the random effects in a generalized linear mixed model (GLMM). Since the random effects are not observed, one approach would be to consider random effects predictions like conditional expectations or modes. However, except for linear mixed models, the distributional properties of such predictions are

unknown. It is thus difficult to judge whether a sample of random effects predictions is consistent with the assumed random effects distribution.

In this paper we pursue instead the following simple idea: consider a pair of random vectors (A, Y) which is assumed to follow some fully specified distribution where A represents the unobserved random effects and Y the observations. If the specified joint model for (A, Y) is correct then the observed data y is a realization from the marginal distribution of Y. Suppose next we generate a simulation A^* from the conditional distribution of A given $Y = y$. Then marginally, A^* and A are identically distributed and correlated. We can thus base goodness-of-fit testing on the sample A^* proceeding just as if A had been observed itself.

The situation becomes more complicated when the joint distribution of (A, Y) depends on some unknown parameter θ. If a proper prior is specified for θ in a Bayesian framework one may consider a simulation (A^*, θ^*) from the posterior of (A, θ) given Y. Again, the distribution of (A^*, θ^*) coincides with the specified distribution of (A, θ) provided the assumed joint model for (θ, A, Y) is correct. In practice, however, one often uses very vague or even improper priors which are not regarded as bona fide components in a joint model for (θ, A, Y).

In this paper we replace θ by a point estimate $\hat{\theta}$. Let $P_{|y,\theta}$ denote the conditional distribution of A given $Y = y$ and indexed by θ. We then base goodness-of-fit tests on a simulation $\tilde{A}$ from $P_{|y,\hat{\theta}}$. In this case $\tilde{A}$ only approximately follows the specified distribution for A due to the effect of replacing the unobserved θ with $\hat{\theta}$. Inspired by Ritz (2004) we derive the asymptotic distribution of the empirical distribution function obtained from the conditional sample $\tilde{A}$ of the random effects. This provides an asymptotic correction for the effect of replacing θ with an estimate.

When the objective is to estimate fixed effects in a GLMM one may argue that the assumptions concerning the shape of the random effects distribution is not critical. However, in many applications, e.g. in quantitative genetics, the random effects themselves and their distributional characteristics are the focal objects of the statistical analysis. A thorough assessment of the goodness-of-fit of the random effects distribution then seems mandatory. Moreover, the approach in this paper is not confined to providing a p-value for a goodness-of-fit test. In the examples in Section 4, exploratory plots of the simulated random effects e.g. disclose patterns of heterogeneity or correlation among the individuals to which the random effects are associated. Such patterns should also be taken into account in an analysis of fixed effects.

In Section 2 we obtain the asymptotic distribution of the empirical distribution function for simulated random effects within the framework of GLMMs with *iid* normal random intercepts. Section 3 is concerned with

the practical implementation of a goodness-of-fit test based on the asymptotic result. Simulation studies and applications are considered in Section 4. Section 5 contains a concluding discussion.

2 Convergence of the empirical distribution function for a conditional sample of random effects

The asymptotic result in this section is derived within the set-up of generalized linear models (GLMs) with *iid* normal random intercepts. Maximum likelihood inference for such models is implemented in e.g. the SAS procedure `nlmixed` or the Stata program `gllamm`.

2.1 Set-up and notation

Denote by $Y = (Y_i)_{i\geq 1}$ a sequence of observation vectors $Y_i = (Y_{i1}, \ldots, Y_{iN_i})$ with associated covariates $X_i = (X_{i1}, \ldots, X_{iN_i})$, $X_{ij} \in \mathbb{R}^p$, $p \geq 1$, and random effects A_i. The (N_i, X_i, A_i, Y_i), $i \geq 1$, are assumed to be independent where N_i is integer valued, (N_i, X_i) follows some unspecified distribution, and given (N_i, X_i), A_i is $N(0, 1)$. The linear predictor for the observation Y_{ij} is $\eta_{ij} = X_{ij}\beta^{\mathsf{T}} + \sigma A_i$ where $\beta \in \mathbb{R}^p$ and $\sigma > 0$. Conditional on $A_i = a_i, N_i = n_i, X_i = x_i$, $Y_{i1}, \ldots, Y_{in_i}$ are conditionally independent with densities of GLM type (see e.g. McCullagh and Nelder, 1989), i.e. the conditional density of Y_{ij} is of the form

$$f(y_{ij}|\psi_{ij}, \phi) = \exp((y_{ij}\psi_{ij} - b(\psi_{ij}))/\phi + c(y_{ij}, \phi))$$

where $\phi > 0$, b and c are certain functions, and $\psi_{ij} = h(\eta_{ij})$ for some one-to-one function h.

The joint distribution of (N_i, X_i, A_i, Y_i), $i \geq 1$, is parametrized by $\theta = (\beta, \sigma, \phi)$ which belongs to an open set $\Theta \subset \mathbb{R}^{p+2}$. We assume that Y is generated under the joint distribution P_{θ_0} corresponding to a specific parameter value $\theta_0 \in \Theta$. Henceforth, probabilities, expectations, and variances are computed with respect to P_{θ_0} unless otherwise stated.

With a slight abuse of notation let $F_{i|y_i,\theta}$ denote the distribution function of A_i given $(N_i, X_i, Y_i) = (n_i, x_i, y_i)$ and let $\hat{\theta}_n$ denote an estimate of θ based on $Y_1, \ldots, Y_n$. For each n, $(\tilde{A}_{1n}, \ldots, \tilde{A}_{nn})$ denotes a sample where $\tilde{A}_{in}$ is generated from $F_{i|Y_i,\hat{\theta}_n}$ and $\tilde{A}_{in}$, $i = 1, \ldots, n$, are independent given Y, $N = (N_i)_i$, and $X = (X_i)_i$. The empirical distribution function based

on $\tilde{A}_{1n}, \ldots, \tilde{A}_{nn}$ is denoted $\tilde{F}_n$. For a finite index set $I \subseteq \mathbb{R}$, the asymptotic distribution of $(\tilde{F}_n(t))_{t\in I}$ is given in the following Section 2.2.

2.2 Asymptotic result

Assume that h and c are continuously differentiable and (for sake of the second result (10) in the Appendix) assume that h' is bounded, and that $|h(A_1)|$, $|b(h(A_1))|$ and $|A_1 b'(h(A_1))|$ have finite expectation. All these assumptions are valid for the common examples of GLMMs considered in Section 4. Assuming in addition that $\hat{\theta}_n$ is asymptotically normal and efficient, we obtain

Theorem 1. *Under the above set-up and assumptions,* $(\tilde{G}_n(t))_{t\in I} = (\sqrt{n}(\tilde{F}_n(t) - \Phi(t)))_{t\in I}$ *converges in distribution to a zero mean Gaussian vector* $(G(t))_{t\in I}$ *with covariances given by*

$$\mathbb{E}G(s)G(t) = \Phi(s \wedge t) - \Phi(s)\Phi(t) - h(s,\theta_0)V(\theta_0)h(t,\theta_0)^{\mathsf{T}},\ s,t \in I \qquad (1)$$

where $h(t,\theta) = \mathbb{E}(\mathrm{d}F_{1|Y_1,\theta}(t)/\mathrm{d}\theta|_{\theta=\theta_0})$, $\Phi(\cdot)$ *is the standard normal distribution function, and* $V(\theta_0)$ *is the asymptotic covariance matrix for* $\hat{\theta}_n$.

Proof. Without loss of generality we can assume $\tilde{A}_{in} = F^{-1}_{i|Y_i,\hat{\theta}_n}(U_i)$ where $U_1, U_2, \ldots$ *iid* uniform on $[0,1]$ and independent of N, X, and Y. Similarly, we let $A^*_i = F^{-1}_{i|Y_i,\theta_0}(U_i)$ and let F^*_n denote the empirical distribution function based on $A^*_1, \ldots, A^*_n$.

We now split $\tilde{G}_n(t)$ as follows:

$$\tilde{G}_n(t) = G^*_n(t) + Z_n(t)$$

where

$$G^*_n(t) = \sqrt{n}(F^*(t) - \Phi(t))$$

and

$$Z_n(t) = \sqrt{n}(\tilde{F}_n(t) - F^*_n(t)).$$

Note that, marginally, the A^*_i are independent standard normal variables. Hence, weak convergence of $(G^*_n(t))_{t\in\mathbb{R}}$ to a zero-mean Gaussian process with covariance function $\Phi(s \wedge t) - \Phi(s)\Phi(t)$ is a classical result, see e.g. Van der Vaart (1998).

Consider now the term $Z_n(t)$ and let $d_{in}(t) = 1[U_i \le F_{i|Y_i,\hat{\theta}_n}(t)] - 1[U_i \le F_{i|Y_i,\theta_0}(t)]$, $m_{in}(t) = \mathbb{E}(d_{in}(t)|N,X,Y) = F_{i|Y_i,\hat{\theta}_n}(t) - F_{i|Y_i,\theta_0}(t)$, and $v_{in}(t) = \mathbb{V}\mathrm{ar}(d_{in}(t) - m_{in}(t)) = \mathbb{E}(|m_{in}(t)|(1 - |m_{in}(t)|))$. Then

$$\sup_{t\in I}\Big(Z_n(t) - \sum_{i=1}^n m_{in}(t)/\sqrt{n}\Big) = \sup_{t\in I}\sum_{i=1}^n (d_{in}(t) - m_{in}(t))/\sqrt{n}. \qquad (2)$$

Note that $\mathbb{C}\mathrm{ov}(d_{in}(t)-m_{in}(t), d_{jn}(t)-m_{jn}(t)) = 0$ when $i \neq j$. Hence, using Chebyshevs inequality,

$$P(\sup_{t\in I} |\sum_{i=1}^n (d_{in}(t)-m_{in}(t))/\sqrt{n}| \geq \epsilon) \leq \sum_{t\in I} P(|\sum_{i=1}^n (d_{in}(t)-m_{in}(t))/\sqrt{n}| \geq \epsilon)$$
$$\leq \mathrm{card}(I) \sup_{t\in I} v_{1n}(t)/\epsilon^2.$$

It follows from (9) in Lemma 1 in the Appendix that the right hand side of (2) is $o_P(1)$ so that $\sup_{t\in I}\left(Z_n(t) - \sum_{i=1}^n m_{in}(t)/\sqrt{n}\right) = o_P(1)$. Moreover, letting $g_i(t,\theta) = \mathrm{d}F_{i|Y_i,\theta}(t)/\mathrm{d}\theta$,

$$\sum_{i=1}^n m_{in}(t)/\sqrt{n} =$$
$$\sqrt{n}(\hat{\theta}_n - \theta_0)\sum_{i=1}^n g_i(t,\theta_0)^\mathsf{T}/n + \sqrt{n}(\hat{\theta}_n - \theta_0)\sum_{i=1}^n (g_i(t,\lambda_{in}) - g_i(t,\theta_0))^\mathsf{T}/n$$

where λ_{in} is between θ_0 and $\hat{\theta}_n$. Note that $\sum_{i=1}^n g_i(t,\theta_0)/n$ converges to $h(t,\theta_0) = \mathbb{E}g_1(t,\theta_0)$ almost surely. Further,

$$P(\sup_{t\in I} |\sum_{i=1}^n (g_i(t,\theta_0) - g_i(t,\lambda_{in}))^\mathsf{T}/n| \geq \epsilon) \leq \mathbb{E}\sup_{t\in I} |g_1(t,\theta_0) - g_1(t,\lambda_{1n})|/\epsilon$$

where by (10) in Lemma 1, $\lim_{n\to\infty} \mathbb{E}\sup_{t\in I} |g_{1,t}(\theta_0) - g_{1,t}(\lambda_{1n})| = 0$ since λ_{1n} tends to θ_0 in probability.

We conclude that

$$\sup_{t\in I} |\tilde{G}_n(t) - G_n^*(t)) - \sqrt{n}(\hat{\theta}_n - \theta_0)h(t,\theta_0)| = o_P(1) \tag{3}$$

and thus $(\tilde{G}_n(t))_{t\in I}$ and $(V_n(t))_{t\in I} = (G_n^*(t) + \sqrt{n}(\hat{\theta}_n - \theta_0)h(t,\theta_0))_{t\in I}$ have the same weak limit. By Pierce (1982), $(V_n(t))_{t\in I}$ and hence $(\tilde{G}_n(t))_{t\in I}$ converges to a zero mean Gaussian vector $(G(t))_{t\in I}$ with covariances $\mathbb{E}G(s)G(t)$ given by (1) for $s,t \in I$. □

Remark 1. Weak convergence of the process $(\tilde{G}_n(t))_{t\in\mathbb{R}}$ essentially follows provided I in (3) can be replaced by $\mathbb{R}$. The main obstacle here is to verify that $\sup_{t\in\mathbb{R}} \sum_{i=1}^n (d_{in}(t) - m_{in}(t))/\sqrt{n}$ is $o_P(1)$. However, convergence of finite dimensional distributions suffices for our application, see Section 3.

3 Implementation of goodness-of-fit tests

In the subsequent simulation studies and applications our goodness-of-fit statistic $T(\tilde{F}_n)$ is a discretized version of the Anderson-Darling statistic, i.e. for any distribution function F,

$$T(F) = \frac{1}{m}\sum_{l=0}^{m-1}\frac{(F(\Phi^{-1}(v_l)) - v_l)^2}{v_l(1-v_l)} \tag{4}$$

where $v_l = (l+0.5)/m$, $l = 0, \ldots, m-1$. As in Ritz (2004) we compute p-values using simulation from the asymptotic distribution of $\tilde{F}_n$ with $h(t, \theta_0)$ and $V(\theta_0)$ replaced by estimates obtained as follows.

Denote by $f_i(\cdot;\theta)$ the density of Y_i given $X_i = x_i$ and $N_i = n_i$, i.e.

$$f_i(y_i;\theta) = \int_{-\infty}^{\infty}\prod_{j=1}^{n_i} f(y_{ij}|\psi_{ij},\phi)\Phi'(a_i)\mathrm{d}a_i, \tag{5}$$

and denote by $u_i(\theta)$ and $j_i(\theta)$ the score function and observed information based on $f_i(Y_i;\theta)$. Let further $\mathbb{E}_\theta$ denote expectation under P_θ. Then

$$h(t,\theta) = \mathbb{E}_\theta \mathrm{d}F_{1|Y_1,\theta}(t)/\mathrm{d}\theta = -\mathbb{E}_\theta u_1(\theta)F_{1|Y_1,\theta}(t)$$

which is approximated by

$$-\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\theta[u_i(\theta)F_{i|Y_i,\theta}(t)|n_i, x_i] \tag{6}$$

where the expectation is conditional on $N_i = n_i$ and $X_i = x_i$. In the case of a linear mixed model, we can calculate the conditional expectations in (6) explicitly. For GLMMs in general, we first compute $u_i(\theta)$ and $f_i(Y_i;\theta)$ using adaptive Gaussian quadrature (see e.g. Pinheiro and Bates, 1995) and compute

$$F_{i|Y_i,\theta}(t) = \int_{-\infty}^{t}\frac{\prod_{j=1}^{N_i} f(Y_{ij}|\psi_{ij},\phi)}{f_i(Y_i;\theta)}\Phi'(a_i)\mathrm{d}a_i = \int_0^{\Phi(t)}\frac{\prod_{j=1}^{N_i} f(Y_{ij}|\psi_{ij},\phi)}{f_i(Y_i;\theta)}\mathrm{d}v_i \tag{7}$$

using a Riemann sum. Secondly, the conditional expectation of $u_i(\theta)F_{i|Y_i,\theta}(t)$ given $N_i = n_i$ and $X_i = x_i$ is in general computed using Monte Carlo simulations of Y_i given $N_i = n_i$ and $X_i = x_i$. In the special case where the Y_{ij} are identically distributed, $j = 1, \ldots, n_i$, we can rewrite the conditional expectation in terms of first a conditional expectation with respect to the scalar

random variable $Y_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij}$ given A_i, N_i, and X_i, and secondly an expectation with respect to A_i. Thereby numerical integration becomes feasible. Similarly, we approximate $V(\theta)$ by the inverse of $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\theta[j_i(\theta)|n_i, x_i]$. For linear mixed models we can again calculate the conditional expectation explicitly while we resort to Monte Carlo or numerical integration for GLMMs in general. Finally, we replace θ_0 by the maximum likelihood estimate $\hat{\theta}_n$.

The simulations $\tilde{A}_{in}$ are in general obtained using adaptive rejection sampling using a t-distribution centered at the mode and scaled by the hessian, respectively, for the conditional density of A_i given (N_i, X_i, Y_i). In the case of linear mixed models, $\tilde{A}_{in}$ can be sampled directly since the conditional distribution of A_i is normal with conditional expectation and variance given explicitly by

$$\mathbb{E}_\theta[A_i|n_i, x_i, y_i] = \frac{\sigma}{\phi/n_i + \sigma^2}\bar{r}_i \text{ and } \mathbb{V}\mathrm{ar}_\theta[A_i|n_i, x_i, y_i] = \frac{1}{1 + n_i\sigma^2/\phi} \tag{8}$$

where $\bar{r}_i = \sum_{j=1}^{n_i}(y_{ij} - x_{ij}\beta^\mathsf{T})/n_i$.

All computations are implemented in `R` (R Development Core Team, 2004) and `c`. Programs can be obtained from the author.

4 Simulation studies and applications

In the first part of this section we compare our approach to the one introduced by Ritz (2004) within the context of linear mixed models. We secondly turn to Poisson-log normal and binomial-logit normal GLMMs.

4.1 Linear mixed models

In the case of linear mixed models, the conditional distribution of A_i given Y_i is normal and known in closed form, see (8). Ritz (2004) base goodness-of-fit testing on the empirical distribution function for standardized estimated conditional expectations $\hat{A}_i(\hat{\theta}_n)/\sigma_i(\hat{\theta}_n)$ where $\hat{A}_i(\theta) = \mathbb{E}_\theta[A_i|N_i, X_i, Y_i]$ and $\sigma_i^2(\theta) = \mathbb{V}\mathrm{ar}_\theta[\hat{A}_i(\theta)|N_i, X_i] = \sigma^2/(\sigma^2 + \phi/N_i)$. We here repeat parts of the simulation study leading to Table 3 in Ritz (2004). Briefly, $n = 50$, $N_i \sim \mathrm{Poisson}(5) + 1$, treatments $0, 10$, or 12 are assigned randomly to units, $\sigma = 1$, and the conditional variance ϕ is 1. The random effects distribution is either $N(0, 1)$, $0.46t(2)$ or $\Gamma(1, 1)$. In addition we consider also $n = 100$.

Table 1 shows Monte Carlo estimates (based on 10000 synthetic data sets) of powers obtained using $T(\tilde{F}_n)$ (see (4)) with $m = 150$, and rejecting when the bootstrap p-value is below the nominal levels $\alpha = 0.01$, $\alpha = 0.05$, or $\alpha = 0.10$, respectively. The numbers in brackets are the powers obtained

n	50			100		
α	0.01	0.05	0.10	0.01	0.05	0.10
N(0,1)	0.02 [0.01]	0.06 [0.06]	0.12 [0.12]	0.02	0.06	0.11
0.46t(2)	0.38 [0.53]	0.47 [0.65]	0.54 [0.72]	0.60	0.70	0.75
$\Gamma(1,1)$	0.34 [0.66]	0.49 [0.83]	0.59 [0.88]	0.64	0.78	0.84

Table 1: Power of goodness-of-fit test $T(\tilde{F}_n)$ with different random effects distributions and nominal levels and with $n = 50$ or $n = 100$. The numbers in brackets are powers obtained in Ritz (2004).

by Ritz (2004). The first row indicates that our test and the Ritz (2004) test are both rather close to the nominal levels. The next two rows show that within the setting of the simulation experiment we need about twice the number of random effects to achieve the power of the Ritz (2004) test. That the Ritz (2004) test is more powerful is not surprising in view of the following considerations: Assuming for a moment that the variance of A_i exists and is equal to one,

$$\mathbb{C}\mathrm{orr}(A_i^*, A_i) = \mathbb{C}\mathrm{orr}(\hat{A}_i, A_i)^2 \text{ where } \mathbb{C}\mathrm{orr}(\hat{A}_i, A_i) = (1 + \phi/(n_i\sigma^2))^{-1/2}.$$

Hence we may expect that A_i is more correlated with $\hat{A}_i(\hat{\theta}_n)$ than with $\tilde{A}_{in}$.

Example 1. To illustrate further we apply our test to the data from Example 2 in Ritz (2004) (602 observations in $n = 56$ groups of roughly the same size, see Damstrup and Nielsen, 2002). We obtain MLEs $\hat{\sigma} = 0.047$ and $\hat{\phi} = 2.840$. With these parameter values, $\mathbb{C}\mathrm{orr}(A_i^*, A_i)$ is smaller than 0.0033. Hence it is an unfavorable situation for our test. Thus Ritz (2004)'s test is highly significant ($p < 0.005$) while ours is not ($p = 0.30$). The example may be considered somewhat extreme since the random effects only contribute with a very small proportion $2.7e - 4$ of the total variance for an observation.

Example 2. As a second example we consider a data set extracted from the Framingham study (see e.g. Zhang and Davidian, 2001). The data set contains repeated measurements of cholestorol level for 384 women of age greater than 50 at study entry. We apply a linear mixed model with individual specific random intercepts and time as explanatory variable. We obtain $\hat{\sigma} = 0.36$ and $\hat{\phi} = 0.24$ and a p-value of 0.004. Figure 1 shows the conditional simulation of the random effects. In accordance with Zhang and Davidian (2001) the plot suggests the presence of a subpopulation of individuals with extraordinary high cholestorol levels.

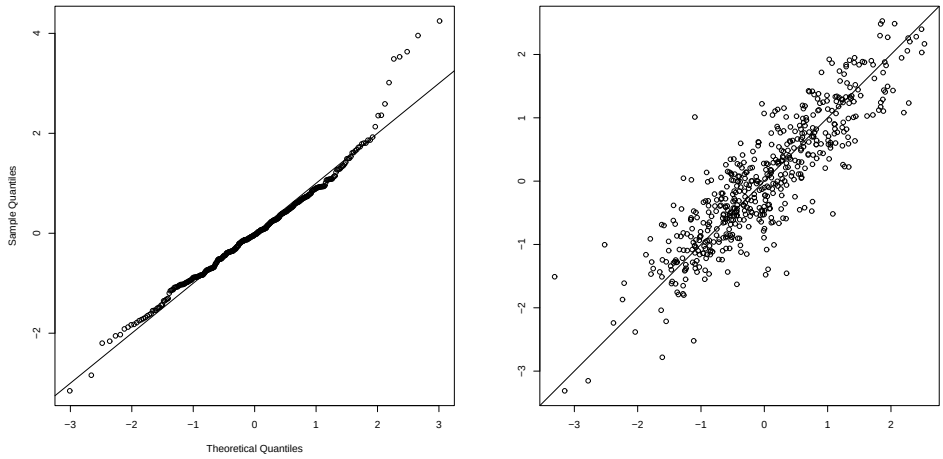

Figure 1: Left: qq-plot for a conditional sample of the random effects for the Framingham data (Example 2) versus the standard normal distribution. Right: a scatterplot of $(\tilde{A}_{in}, \tilde{A}_{(i+1)n})$, $i = 1, \ldots, 533$, for the mumps data (Example 4).

4.2 Poisson regression

We now turn to an example of a non-normal and non-linear model. The observations Y_{ij} are assumed to follow a Poisson-log normal model, i.e. conditional on $A_i = a_i$ and $X_i = x_i$, Y_{ij} is Poisson with expectation $\exp(x_{ij}\beta^{\mathsf{T}} + \sigma a_i)$. In the following simulation study we consider for ease of computation (cf. Section 3) fixed $N_i = 6$ and $X_{ij} = 1$. We use the test statistic $T(\tilde{F}_n)$ with $m = 300$ (the discretization used for the Riemann-approximation of (7)).

Table 2 contains Monte Carlo estimates of the levels for this test for varying values of n, $\mathbb{E}[Y_{ij}]$, and σ when rejecting at the nominal levels 0.01, 0.05 and 0.10. The table indicates that the levels of our test is slightly too

σ	$\mathbb{E}[Y_{ij}]$	$n = 50$			$n = 100$		
0.25	3	0.02	0.06	0.10	0.01	0.05	0.10
0.25	8	0.02	0.06	0.12	0.01	0.06	0.11
0.50	3	0.02	0.07	0.12	0.01	0.05	0.10
0.50	8	0.01	0.07	0.12	0.01	0.06	0.12

Table 2: Levels for the goodness-of-fit statistic $T(\tilde{F}_n)$ when observations are Poisson-log normal with varying parameters and rejecting at the nominal levels 0.01, 0.05, or 0.10.

high for $n = 50$ and fairly close to the nominal levels when $n = 100$ (the

estimated levels are based on 1000 synthetic data sets and are subject to Monte Carlo errors with standard deviations around 0.003, 0.007, and 0.009 for the three nominal levels considered).

The powers in Table 3 are estimated for $0.46t(2)$ and $\Gamma(1,1)$ alternatives and with nominal level 0.05. The power is poor for the small counts case $\sigma = 0.25$ and $\mathbb{E}[Y_{ij}] = 3$ but increases to reasonable values as σ and $\mathbb{E}[Y_{ij}]$ increases. With $N_i = 2$ and with standard normal random effects we obtain

		$n = 50$		$n = 100$	
σ	$\mathbb{E}[Y_{ij}]$	$0.46t(2)$	$\Gamma(1,1)$	$0.46t(2)$	$\Gamma(1,1)$
0.25	3	0.15	0.14	0.22	0.20
0.25	8	0.29	0.34	0.42	0.60
0.50	3	0.32	0.56	0.49	0.84
0.50	8	0.52	0.84	0.75	0.98

Table 3: Powers obtained for non-normal random effects alternatives when $N_i = 6$, observations are conditionally Poisson, and with nominal level 0.05.

levels (not shown) similar to those in Table 2. As might be expected, the powers for the non-normal alternatives decrease when N_i is reduced to 2, see Table 4.

		$n = 50$		$n = 100$	
σ	$\mathbb{E}[Y_{ij}]$	$0.46t(2)$	$\Gamma(1,1)$	$0.46t(2)$	$\Gamma(1,1)$
0.25	3	0.09	0.06	0.13	0.05
0.25	8	0.14	0.12	0.20	0.16
0.50	3	0.17	0.20	0.27	0.34
0.50	8	0.28	0.54	0.50	0.80

Table 4: Powers obtained for non-normal random effects alternatives when $N_i = 2$, observations are conditionally Poisson, and with nominal level 0.05.

Example 3. Our first example of a Poisson regression with random intercept is Model II from Breslow and Clayton (1993) who considered the epileptic seizure data from Thall and Vail (1990). For each of 59 subjects, the observation vector Y_i consists of four counts of epileptic seizures during two week periods prior to each of four clinic visits. The covariates are log baseline count, treatment (placebo or a certain drug), log age, and an indicator for the fourth visit. We obtain $\hat{\sigma} = 0.50$ and a p-value of 0.69. Hence the goodness-of-fit test does not provide evidence against the assumption of normal subject specific random intercepts.

Conditional on the data, our p-value is subject to random variation mainly due to the conditional sampling of the random effects. Figure 2 (left) shows a histogram of p-values obtained for 100 independent repetitions of the goodness-of-fit test applied to the seizure data. The variation of the p-values is considerable but only a small fraction (3%) suggests evidence at the 5% level against the assumption of normal random effects.

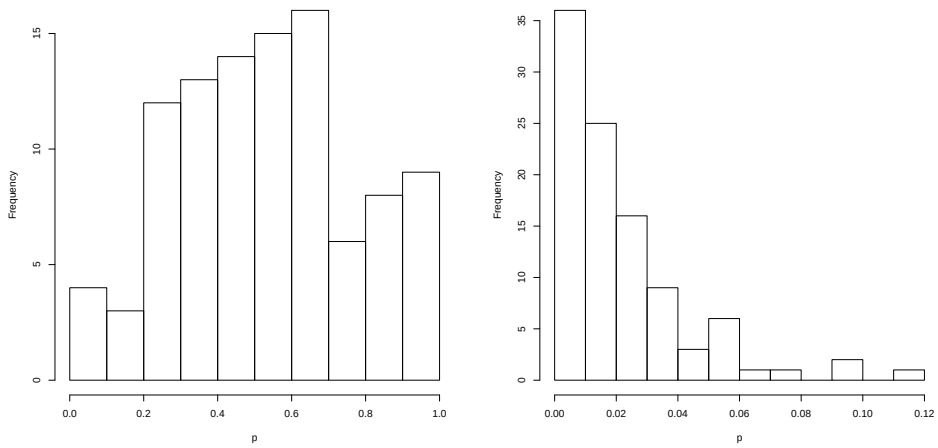

Figure 2: Histograms illustrating the variation of the p-values conditional on (left) the epileptic seizure data (Example 3) and (right) the mumps data (Example 4).

Example 4. The second Poisson regression example is concerned with a time series of 534 monthly counts of mumps cases in New York City, 1928-1972 (Hipel and McLeod, 1994). The data set is obtained from the Time Series Data Library maintained by Rob Hyndman and Muhammed Akram. There is a pronounced seasonal variation so we include month (1-12) as a categorical covariate. In addition, we include a quadratic term in time (measured in units of months) to take into account large scale variation. A random intercept for each observation accounts for possible overdispersion. We obtain $\hat{\sigma} = 0.42$ (with approximate confidence interval $[0.39, 0.45]$) and $p = 0.01$ so there is evidence of overdispersion and evidence against the assumption of independent normal random effects. Especially the assumption of independence seems dubious in this example and the right scatterplot of $(\tilde{A}_{in}, \tilde{A}_{(i+1)n})$ in Figure 1 strongly indicates serial correlation.

As in the previous example, Figure 2 (right) shows the distribution of p-values obtained from 100 independent repetitions of the goodness-of-fit test applied to the mumps data. The distribution of the p-values is concentrated on small values - 89% of the p-values fall below the 5% significance level.

4.3 Logistic regression

We briefly comment on the case of a logistic regression with random effects. The observations Y_{ij} are conditionally Bernouilli with $\text{logit}(\mathbb{E}[Y_{ij}|x_i, a_i]) = \beta^{\mathsf{T}} + \sigma a_i$ and we consider fixed $N_i = 6$, $n = 50$ or 100, $\beta = 0$ or 1, and $\sigma = 0.5$ or 1. In the case of standard normal random effects the levels (not shown) are close to the nominal levels. For the non-normal alternatives $\Gamma(1, 1)$ and $0.46t(2)$ we obtain very poor powers (not shown) only slightly bigger than the nominal level. For a logistic regression, each observation Y_{ij} only contains very limited information on A_i, so very large values of N_i or n are needed to get reasonable powers.

5 Discussion

Our goodness-of-fit test is targeted at the random effects but a rejection could be due to a wrongly specified conditional distribution of $Y|A$. So the test should of course be accompanied by an assessment of the conditional distribution of the observations given the random effects. In this connection it might be helpful to consider "simulated residuals" obtained by replacing A_i by $\tilde{A}_{in}$ in the linear predictor for Y_{ij}.

An alternative to the use of the asymptotic result in Theorem 1 is a parametric bootstrap where $T(\tilde{F}_n)$ is compared with $T(\tilde{F}_n^b)$, $b = 1, \ldots, B$, where $\tilde{F}_n^b$ is obtained as $\tilde{F}_n$ but from synthetic data simulated under $P_{\hat{\theta}_n}$. In the data examples, similar p-values were obtained with the parametric bootstrap and the asymptotic approach, but the parameter estimation for each bootstrap replicate data set is time consuming. The parametric bootstrap on the other hand allows for consideration of goodness-of-fit statistics for which the asymptotic distribution is not known.

Our statistic $T(\tilde{F}_n)$ may be viewed as an approximate simulated replicate of the unobserved statistic $T(F_n)$ given by (4) with F equal to the empirical distribution function F_n for $A_1, \ldots, A_n$. We then compare $T(\tilde{F}_n)$ with the asymptotic sampling distribution under the assumed joint model for (N_i, X_i, A_i, Y_i), $i \geq 1$. In a Bayesian framework, one might instead following Dey *et al.* (1998) consider the entire posterior distribution $\mathcal{D}(T(F_n)|y)$ of $T(F_n)$ (or any other goodness-of-fit statistic) given $(Y_1, \ldots, Y_n) = y$ and compare this with its sampling distribution (i.e. the sampling distribution of the data when the unknown parameters are sampled from the priors). The approach in Dey *et al.* (1998) relies on the use of informative priors and is computationally demanding due to the need for computing posterior distributions for replicate data sets.

Acknowledgment I thank Christian Ritz for helpful discussions and assistance with data.

References

Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Am. Statist. Assoc.* **88**, 9–25.

Damstrup, M. L. & Nielsen, M. M. (2002). Fytaseaktivitet under rugbrødsfremstilling. *Course report*, Royal Veterinary and Agricultural University, Copenhagen.

Dey, D. K., Gelfand, A. E., Swartz, T. B. & Vlachos, P. K. (1998). A simulation-intensive approach for checking hierarchical models. *Test* **7**, 325–346.

Hipel, K. W. & McLeod, A. I. (1994). *Time series modelling of water resources and environmental systems*. Elsevier.

McCullagh, P. & Nelder, J. A. (1989). *Generalized linear models*. Chapman and Hall, London, 2nd edition.

Pierce, D. A. (1982). The asymptotic effect of substituting estimators for parameters in certain types of statistics. *Annals of Statistics* **10**, 475–478.

Pinheiro, J. C. & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics* **4**, 12–35.

R Development Core Team (2004). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Ritz, C. (2004). Goodness-of-fit tests for mixed models. *Scandinavian Journal of Statistics* **31**, 443–458.

Thall, P. F. & Vail, S. C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics* **46**, 657–671.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Zhang, D. & Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics* **57**, 795–802.

Address for correspondence: Rasmus Waagepetersen, Institute of Mathematical Sciences, Aalborg University, Fredrik Bajersvej 7G, DK-9220 Aalborg. email: `rw@math.aau.dk`

A A technical lemma

The set-up, notation, and assumptions are as in Section 2.

Lemma 1. *Assume that θ_n is a sequence which tends to θ_0 in probability. Then*

$$\lim_{n\to\infty} \mathbb{E} \sup_{t\in\mathbb{R}} |F_{1|Y_1,\theta_n}(t) - F_{1|Y_1,\theta_0}(t)| = 0 \tag{9}$$

and

$$\lim_{n\to\infty} \mathbb{E} \sup_{t\in\mathbb{R}} |g_1(t,\theta_n) - g_1(t,\theta_0)| = 0 \tag{10}$$

where $g_1(t,\theta) = \mathrm{d}F_{1|Y_1,\theta}(t)/\mathrm{d}\theta$.

Proof. Recall (5) and (7). It is a standard fact for GLMs that

$$f(y_{1j}|\psi_{1j},\phi) < \tilde{f}(y_{1j}|y_{1j},\phi) \tag{11}$$

where $\tilde{f}(y_{1j}|\mu_{1j},\phi) = f(y_{1j}|(b')^{-1}(\mu_{1j}),\phi)$ is obtained by reparametrization in terms of the mean parameter $\mu_{1j} = b'(\psi_{ij})$. Since $\tilde{f}(y_{1j}|y_{1j},\phi)$ does not depend on a_1 it follows using dominated convergence that $f_1(y_1;\theta)$ and $F_{1|y_1;\theta}(t)$ are continuous functions of (n_1,x_1,y_1,θ) and (n_1,x_1,y_1,θ,t), respectively.

Consider now (9). Given $\epsilon > 0$ we establish in the following that there is a n_1 so that

$$\mathbb{E} \sup_{t\in\mathbb{R}} |F_{1|Y_1,\theta_n}(t) - F_{1|Y_1,\theta_0}(t)| < \epsilon \tag{12}$$

when $n \geq n_1$. Choose a closed ball B_0 centered at θ_0 and of positive radius and a compact set C such that $P((N_1,X_1,Y_1) \in C, \theta_n \in B_0) \geq 1-\epsilon/4$ whenever $n \geq n_0$ for some sufficiently large n_0. Using (11) and continuity,

$$\sup_{(n_1,x_1,y_1,\theta)\in C\times B_0} F_{1|y_1,\theta}(t) < M\Phi(t)$$

for some $0 < M < \infty$ whereby

$$\lim_{t\to-\infty} \sup_{(n_1,x_1,y_1,\theta)\in C\times B_0} F_{1|y_1,\theta}(t) \to 0$$

and

$$\lim_{t\to\infty} \sup_{(n_1,x_1,y_1,\theta)\in C\times B_0} 1 - F_{1|y_1,\theta}(t) \to 0.$$

Consider $n \geq n_0$. We can now choose t_1 and t_2 so that

$$\mathbb{E} \sup_{t<t_1 \text{ or } t>t_2} |F_{1|Y_1,\theta_n}(t) - F_{1|Y_1,\theta_0}(t)| <$$
$$\epsilon/4 + \mathbb{E} \sup_{t<t_1 \text{ or } t>t_2} 1[(N_1, X_1, Y_1) \in C, \theta_n \in B_0)]|F_{1|Y_1,\theta_n}(t) - F_{1|Y_1,\theta_0}(t)| < \epsilon/2.$$

Moreover, for all $\eta > 0$ we can choose $n_1 \geq n_0$ so that $P((N_1, X_1, Y_1) \in C, |\theta_n - \theta_0| > \eta) < \epsilon/4$ whenever $n \geq n_1$. Then

$$\mathbb{E} \sup_{t\in[t_1,t_2]} |F_{1|Y_1,\theta_n}(t) - F_{1|Y_1,\theta_0}(t)| <$$
$$\epsilon/4 + \mathbb{E} \sup_{t\in[t_1,t_2]} 1[(N_1, X_1, Y_1) \in C, |\theta_n - \theta_0| \leq \eta]|F_{1|Y_1,\theta_n}(t) - F_{1|Y_1,\theta_0}(t)|.$$

Since $F_{1|y_1,\theta}(t)$ is uniformly continuous on $C \times B_0 \times [t_1, t_2]$,

$$\mathbb{E} \sup_{t\in[t_1,t_2]} 1[(N_i, X_i, Y_i) \in C, |\theta_n - \theta_0| \leq \eta]|F_{1|Y_1,\theta_n}(t) - F_{1|Y_1,\theta_0}(t)| < \epsilon/4$$

provided η is small enough and (12) follows.

Regarding (10), one can check that for any $\tilde{\theta} \in \Theta$ there exists a neighbourhood U around $\tilde{\theta}$ so that

$$\frac{\partial}{\partial \theta_l} \prod_{j=1}^{n_1} f(y_{1j}|\psi_{1j}, \phi) < M(a_1), l = 1, \ldots, p+2,$$

for $\theta \in U$ where $\int M(a_1)\Phi'(a_1)\mathrm{d}a_1 < \infty$. It follows that $f_1(y_1; \theta)$ and $F_{1|Y_1,\theta}(t)$ are continuously differentiable with respect to θ and that interchange of the order of differentiation and integration is allowed. Thus (10) can be handled along the same lines as (9). □