Review and comparison of measures of explained variation and model selection in linear mixed-effects models

Cantoni, Eva; Jacot, Nadège; Ghisletta, Paolo

How to cite

CANTONI, Eva, JACOT, Nadège, GHISLETTA, Paolo. Review and comparison of measures of explained variation and model selection in linear mixed-effects models. In: Econometrics and statistics, 2024, vol. 29, p. 150–168. doi: 10.1016/j.ecosta.2021.05.005

This publication URL: https://archive-ouverte.unige.ch/unige:153133
Publication DOI: 10.1016/j.ecosta.2021.05.005

Econometrics and Statistics 29 (2024) 150–168

Contents lists available at ScienceDirect

Econometrics and Statistics

journal homepage: www.elsevier.com/locate/ecosta

Review and comparison of measures of explained variation and model selection in linear mixed-effects models

Eva Cantoni[a,*], Nadège Jacot[b], Paolo Ghisletta[c]

[a] Research Center for Statistics and Geneva School of Economics and Management, University of Geneva, Switzerland
[b] Research Center for Statistics, Geneva School of Economics and Management and Faculty of Psychology and Educational Sciences, University of Geneva, Switzerland
[c] Faculty of Psychology and Educational Sciences, University of Geneva, Faculty of Psychology, Swiss Distance University Institute, Brig, and Swiss National Centre of Competence in Research LIVES, University of Geneva, Switzerland

ARTICLE INFO

Article history:
Received 1 November 2020
Revised 20 May 2021
Accepted 25 May 2021
Available online 2 June 2021

Keywords:
Linear mixed-effects model
Explained variation
Model adequacy
Model selection

ABSTRACT

In linear mixed-effects models, several frequentist and Bayesian measures have been proposed to evaluate model adequacy or/and to perform model selection. First, a large set of these measures are selected, presented with comparable notations, discussed in their strengths, weaknesses, and applicability range, and finally commented upon regarding their limitations. Then, these measures are illustrated on the home radon levels data (Gelman & Pardoe, Technometrics, 241-251, 48, 2006). Next, an extensive simulation study is carried out, to evaluate their sensitivity in selecting the correct model from a series of simpler models containing fewer parameters. Finally, recommendations on the use of these different measures are provided.[1]

1. Introduction

The linear mixed-effects model (LMM; a.k.a. linear multilevel model, hierarchical linear model, or random-effects model) is widely used, especially to analyze clustered data. Whether in educational sciences, psychology, medicine, biology or other domains, researchers need tools to compare alternative models and to evaluate their adequacy with respect to the selected data (e.g., Leschinski and Sibbertsen, 2019; Kiviet, 2020; Ko and Hjort, 2019). Several measures, both in the frequentist and in the Bayesian framework, are used for this purpose and have different characteristics (e.g., Gruber and West, 2017). Some of them allow for performing selection of alternative models that differ in their fixed and/or random effects, others allow for assessing the adequacy of a given model, while some allow for both selecting among alternatives and evaluating their adequacy.

For model selection, the most frequently used measures are the Akaike Information Criterion (AIC; Akaike, 1974), the Bayesian Information Criterion (BIC; Schwarz, 1978) and the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002). The Likelihood Ratio Test (LRT; Wilks, 1938) can also be used, but only to compare two nested models (i.e., the parameters

* Corresponding author. Tel.: +41223798240.
E-mail addresses: Eva.Cantoni@unige.ch (E. Cantoni), nadege.jacot4@gmail.com (N. Jacot), Paolo.Ghisletta@unige.ch (P. Ghisletta).

[1] Additional results are available in the Supplementary Material.

https://doi.org/10.1016/j.ecosta.2021.05.005

of the more parsimonious model constitute a subset of the parameters of the larger model). We thus will not discuss the LRT here. Vaida and Blanchard (2005) proposed a conditional AIC for LMMs, which was then generalized by Liang et al. (2008) and Greven and Kneib (2010) (cf. Section 3.2.1). Pu and Niu (2006) extended the Generalized Information Criterion (GIC) from linear regression to LMMs. In the Bayesian framework, Wheeler et al. (2010) proposed a partitioned DIC to assess local model fits instead of a single DIC value for the entire model. In the same perspective of model selection, Jiang et al. (2008) introduced a class of strategies called fence methods for LMMs and generalized LMMs. Other authors, such as (Orelien and Edwards, 2008) and (Edwards et al., 2008), interested in R^2 measures (estimates of variance of the dependent variable explained by the independent variables, Draper and Smith, 1998), focused on the selection of the fixed effects. Moreover, Orelien and Edwards (2008) showed that measures computed by setting the predicted random effects to zero (called marginal measures) are appropriate for that purpose. Chen and Dunson (2003) developed methods to select random effects only. Bondell et al. (2010) used a modified Cholesky decomposition in order to simultaneously select fixed and random effects. For a review and comparison of model selection strategies, we refer to (Müller et al., 2013).

For evaluating model adequacy, several papers considered extensions of the classical R^2 measure due to the simplicity of interpretation. Unfortunately, few of the available measures are absolute in the sense that they can be interpreted without referencing to a comparison (often called null) model. At the opposite, a relative measure can only be interpreted in comparison with another model. This is the case of the measures of model selection defined above and also of most of the definitions of R^2 that require the specification of a null model. Among the measures that assess model adequacy, Snijders and Bosker (1994), Xu (2003), Liu et al. (2008), Gelman and Pardoe (2006) and Nakagawa and Schielzeth (2013) presented extensions of the R^2 measure (cf. Section 3). For generalized LMMs, Zheng (2000), Nakagawa and Schielzeth (2013), and Nakagawa et al. (2017) extended some goodness-of-fit (GOF) measures of a generalized linear model, and Pan and Lin (2005) developed graphical and numerical methods. Finally for nonlinear mixed-effects models, Vonesh et al. (1996) and Vonesh and Chinchilli (1996) proposed a marginal and a conditional version of a concordance correlation coefficient, in addition to a measure of explained residual variation, respectively. We note that in the literature, extensions of R^2 are sometimes presented as GOF measures as in Vonesh and Chinchilli (1996) or (Liu et al., 2008) but, as (Korn and Simon, 1991) highlighted, it is important to distinguish measures of explained variation from GOF. In particular, although some authors use R^2 for model selection (e.g., Xu, 2003), this measure cannot decrease, and usually increases with the addition of predictors to the model. Thus, to use R^2 for model selection, a penalty function (cf. Section 3) is necessary to account for the increase in model complexity (i.e., loss in model parsimony).

In this paper, we are particularly interested in extensions of R^2 and in information criteria. These measures can be used together, with the latter used to compare models and the former used to evaluate the overall quality of the selected model. The considered measures, and their characteristics, are listed in Table 1.

The first column of Table 1 specifies whether a measure can be used to evaluate model adequacy due to both fixed and random effects (A), or to evaluate model adequacy due to fixed effects (F), or to perform model selection (S). Some of the considered measures have dual use as they can be used to both evaluate model adequacy and perform model selection (categories A&S and F&S). In the following, when speaking about overall model adequacy, we imply model adequacy due to both fixed and random effects. Column 2 shows if the measure is absolute or relative and column 3 indicates which relative measure requires the specification of a null model. Finally, the last column concerns the interpretation, as a measure of explained variation, or as a concordance coefficient between observed and predicted values, or as an information criterion, or differently. For instance, D_{rand} is a measure of the proportional reduction in deviance.

The aim of this paper is to compare the measures considered in Table 1. In particular, we compare measures belonging to (a) categories A and A&S; (b) category F&S; and (c) category S. In order to do so, we conduct a simulation study using the home radon levels data of (Gelman and Pardoe, 2006), which have initially motivated our work. In our simulation study, we manipulate five parameters of a random intercept and random slope model. We thus identify which of the considered measures are the most sensitive to these modifications and, among those allowing for model selection, which ones identify the correct model among a series of seven nested alternatives.

In Section 2 we define the LMM and explicit notation. In Section 3 we present and discuss the considered measures. We present and analyze the home radon levels data over which we base the simulation study in Section 4. Then, we describe the simulation study, and present and comment the results in Section 5. Finally we discuss the results, where it appears that P_{rand} and the conditional $r_{c,a}$ are useful to check the overall adequacy of the model at hand; the marginal versions of P_{rand} and $r_{c,a}$ are the most promising measures, among those investigated here, to identify the best set of fixed effects; and the conditional AIC fares best with REML estimation to compare LMMs.

2. Linear mixed-effects model (LMM)

Assume the following LMM (e.g., Laird and Ware, 1982; Bryk and Raudenbush, 1992; Skrondal and Rabe-Hesketh, 2004; Goldstein, 2011):

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, m, \tag{1}$$

where $\mathbf{y}_i = [y_{i1} \cdots y_{in_i}]'$ is the $n_i \times 1$ vector of responses for group i, $\mathbf{X}_i$ is the $n_i \times p$ design matrix for fixed effects for group i, $\boldsymbol{\beta}$ is the $p \times 1$ vector of unknown fixed effects parameters, $\mathbf{Z}_i$ is the $n_i \times q$ design matrix for random effects for group i, $\mathbf{b}_i$ is the $q \times 1$ vector of unobservable random effects for group i and $\boldsymbol{\epsilon}_i$ is the $n_i \times 1$ vector of errors. We assume $\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$,

Table 1
Characteristics of the considered measures. A=overall model adequacy; F=model adequacy due to fixed effects; S=model selection; Abs=absolute; Rel=relative; EV=explained variation; C=concordance coefficient between observed and predicted values; IC=information criterion. m.=marginal; c.=conditional.

		Category	Type	Null model	Measure of
Gelman and Pardoe (2006)	$R^2_{\text{"level"}}$	A	Abs		EV
	$\lambda_{\text{"level"}}$	A	Abs		other
Zheng (2000)	D_{rand}	A	Rel	required	other
	c	A	Abs		C
Xu (2003)	R^2_X	A	Rel	required	EV
	ρ^2_X	A	Rel	required	EV
Liu et al. (2008)	R^2_{T}	A	Rel	required	EV
Vonesh et al. (1996)	c. $r_{c,a}$	A&S	Abs		C
Vonesh and Chinchilli (1996)	c. $R^2_{VC,a}$	A&S	Rel	required	EV
Zheng (2000)	P_{rand}	A&S	Rel	required	other
Xu (2003)	r^2_X	A&S	Rel	required	EV
Liu et al. (2008)	$R^2_{\text{TF},a}$	A&S	Rel	required	EV
Snijders and Bosker (1994)	R^2_1	F&S	Rel	required	EV
	R^2_2	F&S	Rel	required	EV
Vonesh et al. (1996)	m. $r_{c,a}$	F&S	Abs		C
Vonesh and Chinchilli (1996)	m. $R^2_{VC,a}$	F&S	Rel	required	EV
Zheng (2000)	m. D_{rand}	F&S	Rel	required	other
	m. P_{rand}	F&S	Rel	required	other
	m. c	F&S	Abs		C
Xu (2003)	m. R^2_X	F&S	Rel	required	EV
Liu et al. (2008)	$R^2_{\text{F},a}$	F&S	Rel	required	EV
Akaike (1974)	mAIC	S	Rel		IC
Schwarz (1978)	BIC	S	Rel		IC
Vaida and Blanchard (2005)	cAIC	S	Rel		IC
Spiegelhalter et al. (2002)	DIC	S	Rel		IC

where $\mathcal{N}$ is a Gaussian distribution with mean $\mathbf{0}$ and $q \times q$ covariance matrix $\mathbf{D}$ and $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{n_i})$, with $\mathbf{I}_{n_i}$ the $n_i \times n_i$ identity matrix. The variance of the response $\mathbf{y}_i$ is thus $\boldsymbol{\Sigma}_i = \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i' + \sigma^2 \mathbf{I}_{n_i}$. The total number of observations is $N = \sum_{i=1}^{m} n_i$.

The $n_i \times 1$ vector of predicted values for group i is $\hat{\mathbf{y}}_i = \mathbf{X}_i \hat{\boldsymbol{\beta}} + \mathbf{Z}_i \hat{\mathbf{b}}_i$, where $\hat{\boldsymbol{\beta}}$ is the vector of estimated fixed effects and $\hat{\mathbf{b}}_i$ are the random effects predictions. In the frequentist framework, the parameters of $\boldsymbol{\Sigma}_i$ can be estimated with maximum likelihood (ML), or with restricted ML (REML) and plugged into $\hat{\boldsymbol{\beta}}$, and the random effects predictions are the conditional modes (best linear unbiased predictors, or BLUPs). For the Bayesian estimation, it is standard to consider noninformative normal priors for the coefficients associated with the fixed effects. The appropriate choice of priors for the parameters of the covariance matrix $\mathbf{D}$ is less obvious and there is still a debate among specialists. In this article, we follow (Gelman and Pardoe, 2006) and thus use noninformative uniform priors for the parameters of the covariance matrix $\mathbf{D}$.

Given its popularity, we present the LMM with two levels (e.g., children within classrooms), which is consistent with the reviewed literature. Model (1) above is in this case the LMM as described in Laird and Ware (1982). We observe in Section 3 that some of the considered measures can be applied to two-level LMMs (Snijders and Bosker, 1994), while others can be applied to nonlinear mixed-effects models (e.g., Vonesh et al., 1996).

3. Measures

In this Section, we present the considered measures, with a particular effort to make the various notations comparable. Some measures require the specification of a null model, which is either one containing only a fixed intercept or one with a fixed intercept and a random intercept. For subsequent use, we define $\bar{y}$, the grand mean of the observed values y_{ij}, $\mathbf{1}_{n_i}$,

the $n_i \times 1$ unit vector, $\mathbf{I}_{n_i}$, the $n_i \times n_i$ identity matrix, $\hat{y}$, the grand mean of the predicted values $\hat{y}_{ij}$, $\hat{\mathbf{y}}_{i0}$, the $n_i \times 1$ vector of fitted values for group i obtained with the null model and $\sigma_0^2 \mathbf{I}_{n_i}$, the covariance matrix of the errors of the null model.

3.1. Measures of model adequacy only and of model adequacy and model selection

We introduce and compare here the measures belonging to categories A, A&S and F&S (cf. Table 1).

3.1.1. Gelman and Pardoe (2006)

For a LMM with L variance components, Gelman and Pardoe (2006) presented two Bayesian measures that summarize information in the data at each "level" of the model. We write level in quotation marks because it corresponds to the separate variance components, rather than to the more usual definition based on the hierarchy of the data. For instance, consider this varying-intercept model:

$$\begin{array}{ll} y_{ij} & = \beta_0 + b_{i0} + \beta_1 x_{ij} + \epsilon_{ij}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n_i \\ b_{i0} & \sim \mathcal{N}(0, \tau^2), \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \end{array} \tag{2}$$

with a predictor x_{ij}. Model (2) can be written hierarchically with $\beta_{0i} = \beta_0 + b_{i0}$ as

$$y_{ij} \sim \mathcal{N}(\beta_{0i} + \beta_1 x_{ij}, \sigma^2), \quad \beta_{0i} \sim \mathcal{N}(\beta_0, \tau^2).$$

This model has two "levels" (data y_{ij}, intercepts β_{0i}) with a different variance component at each "level" (σ^2, τ^2). The model is defined at each "level" $l = 1, \ldots, L$ as

$$\zeta_k^{(l)} = \nu_k^{(l)} + e_k^{(l)},$$

for $k = 1, \ldots, K^{(l)}$. $\zeta_k^{(1)}$ corresponds to y_{ij} in (1) and for $l > 1$, $\zeta_k^{(l)}$ are the random coefficients β_{0i}, so that $K^{(l)} = 2$ in this example. $\nu_k^{(l)}$ are the linear predictors and $e_k^{(l)}$ are the errors that follow a distribution with mean 0 and standard deviation $\sigma^{(l)}$. For instance, for model (2) and for $l = 1$, we have $\zeta_k^{(1)} = y_{ij}$, $\nu_k^{(1)} = \beta_0 + b_{0i} + \beta_1 x_{ij}$, $e_k^{(1)} = \epsilon_{ij}$ and $\sigma^{(1)} = \sigma$. For $l = 2$, we have $\zeta_k^{(2)} = \beta_{0i}$, $\nu_k^{(2)} = \beta_0$, $e_k^{(2)} = b_{i0}$ and $\sigma^{(2)} = \tau$.

Subsequently, we suppress the superscripts (l), as in Gelman and Pardoe (2006), because we work with each "level" separately. The variation explained by the linear predictors ν_k for each "level" is defined in the population by $1 - [\mathrm{E}(\mathrm{Var}(e_k))][\mathrm{E}(\mathrm{Var}(\zeta_k))]^{-1}$ and is computed as

$$R^2_{\text{"level"}} = 1 - \frac{\mathrm{E}\left(\mathbf{V}_{k=1}^{K}(\hat{e}_k)\right)}{\mathrm{E}\left(\mathbf{V}_{k=1}^{K}(\hat{\zeta}_k)\right)},$$

where "E" is the posterior mean, "Var" is the posterior variance, "$\mathbf{V}$" is the finite-sample variance operator ($\mathbf{V}_{i=1}^{m}(x_i) = (m-1)^{-1} \sum_{i=1}^{m} (x_i - \bar{x})^2$), and $\hat{e}_k$ and $\hat{\zeta}_k$ are the estimates of e_k and ζ_k, respectively. The expectations are estimated by averaging over posterior simulation draws, which gives rise to a "Bayesian adjusted R^2", that is a generalization of the classical adjusted R^2 in regression. $R^2_{\text{"level"}}$ usually takes values between 0 and 1. For each "level," the values of 0 and 1 indicate, respectively, a poor and a perfect fit with respect to the error variance explained at each "level." If $R^2_{\text{"level"}}$ is negative, the prediction is so poor that the estimated error variance is larger than the variance of the data.

The measure that summarizes the average amount of pooling at each "level" is the pooling factor $\lambda_{\text{"level"}}$, which is defined in the population by $1 - [\mathrm{Var}(\mathrm{E}(e_k))][\mathrm{E}(\mathrm{Var}(e_k))]^{-1}$ and is computed as

$$\lambda_{\text{"level"}} = 1 - \frac{\mathbf{V}_{k=1}^{K}\left(\mathrm{E}(\hat{e}_k)\right)}{\mathrm{E}\left(\mathbf{V}_{k=1}^{K}(\hat{e}_k)\right)}.$$

This measure ranges from 0 to 1, where 0 and 1 correspond to no and, respectively, complete pooling. For model (2), the complete-pooling model is $y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}$ with common estimates $\forall i$ and the no-pooling model is $y_{ij} = \beta_{0i} + \beta_1 x_{ij} + \epsilon_{ij}$ with the m β_{0i}'s estimated by least squares. A low pooling factor $\lambda_{\text{"level"}} < 0.5$ indicates a higher degree of within-group information than population-"level" information. A high pooling factor $\lambda_{\text{"level"}} > 0.5$ indicates a higher degree of population-"level" information than within-group information.

3.1.2. Snijders and Bosker (1994)

Two measures of modeled variation, one at each level of a two-level LMM, are defined. This model is equivalent to model (1) in which levels 1 and 2 correspond to the subject level j and group level i, respectively, and $\mathbf{D} = \tau^2 \mathbf{I}_q$. These measures are defined for two-level models only and they require a null model, which contains a fixed intercept and a random intercept, with variance τ_0^2, and for which $\sigma^2 = \sigma_0^2$.

The level-1 modeled proportion of variation is defined in the population as the proportional reduction in mean squared prediction error for $y_{i\cdot}$, $1 - \left(\mathrm{var}(y_{i\cdot} - \mathbf{X}_{i\cdot}\boldsymbol{\beta})\right)(\mathrm{var}(y_{i\cdot}))^{-1}$. The corresponding criterion is

$$R_1^2 = 1 - \frac{\widehat{\mathrm{var}}(y_{ij} - \mathbf{X}_{ij}\boldsymbol{\beta})}{\widehat{\mathrm{var}}(y_{ij})},$$

where $\widehat{\text{var}}$ is obtained by plugging-in the parameter estimates in the population formula and $\mathbf{X}_{ij}$ is the $1 \times p$ vector of fixed effects for subject j in group i.

The level-2 modeled proportion of variation is defined in the population as the proportional reduction in mean squared prediction error for $\bar{y}_{i.}$, $1 - \left(\text{var}(\bar{y}_{i.} - \overline{\mathbf{X}}_{i.}\boldsymbol{\beta})\right)(\text{var}(\bar{y}_{i.}))^{-1}$. The corresponding criterion is

$$R_2^2 = 1 - \frac{\widehat{\text{var}}(\bar{y}_{i.} - \overline{\mathbf{X}}_{i.}\boldsymbol{\beta})}{\widehat{\text{var}}(\bar{y}_{i.})},$$

where $\bar{y}_{i.}$ and $\overline{\mathbf{X}}_{i.}$ are the group means of y_{ij} and $\mathbf{X}_{ij}$, respectively.

For R_1^2 and for R_2^2 for balanced data ($n_i = n\ \forall i$), the numerator is the sample variance of the model of interest and the denominator is the sample variance of the null model. For example, for model (2), the criteria that estimate the population parameters are $R_1^2 = 1 - (\hat{\sigma}^2 + \hat{\tau}^2)/(\hat{\sigma}_0^2 + \hat{\tau}_0^2)$ and $R_2^2 = 1 - (\hat{\sigma}^2/n + \hat{\tau}^2)/(\hat{\sigma}_0^2/n + \hat{\tau}_0^2)$, where $\hat{\sigma}^2$, $\hat{\tau}^2$, $\hat{\sigma}_0^2$ and $\hat{\tau}_0^2$ are the estimates of σ^2, τ^2, σ_0^2 and τ_0^2 respectively.

In the case of unbalanced data, the authors advise to use a representative value of n_i, such as the harmonic mean $\left(m^{-1}\sum_i n_i^{-1}\right)^{-1}$. The interpretation of the R_1^2 and R_2^2 is the same as the traditional coefficient of determination. R_1^2 and R_2^2 identify the proportion of variance explained in y_{ij} and $\bar{y}_{i.}$, respectively. Population values lie between 0 and 1. Negative values are possible when the fixed part of the model is misspecified.

3.1.3. Vonesh et al. (1996)

The model concordance correlation coefficient for generalized nonlinear mixed-effects models is defined as

$$r_c = 1 - \frac{\sum_{i=1}^m (\mathbf{y}_i - \hat{\mathbf{y}}_i)'(\mathbf{y}_i - \hat{\mathbf{y}}_i)}{\sum_{i=1}^m (\mathbf{y}_i - \bar{y}\mathbf{1}_{n_i})'(\mathbf{y}_i - \bar{y}\mathbf{1}_{n_i}) + \sum_{i=1}^m (\hat{\mathbf{y}}_i - \hat{y}\mathbf{1}_{n_i})'(\hat{\mathbf{y}}_i - \hat{y}\mathbf{1}_{n_i}) + N(\bar{y} - \hat{y})^2}.$$

Initially introduced by Lin (1989) to measure the degree of agreement between pairs of observations, r_c is interpretable as a concordance correlation coefficient between observed and predicted values.

To assess the GOF associated with the fixed effects and to select the best set of fixed effects, a marginal model concordance correlation is obtained by setting $\hat{\mathbf{b}}_i = 0$ in $\hat{\mathbf{y}}_i$. If $\hat{\mathbf{b}}_i$ are not set to zero, as in the original definition, r_c is referred to as the conditional model concordance correlation and it assesses the GOF associated with fixed and random effects. The range of values of r_c is between -1 and 1 as the usual Pearson correlation but with a slightly different interpretation. Indeed, r_c measures the level of agreement, or concordance, between $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$, and a value of 1 indicates perfect fit, while a value smaller than or equal to zero indicates lack of fit. Adjusted values for the number of parameters in $\boldsymbol{\beta}$ are defined as $r_{c,a} = 1 - N(N-p)^{-1}(1 - r_c)$, which allows using the conditional $r_{c,a}$ for model selection.

3.1.4. Vonesh and Chinchilli (1996)

Another measure of explained residual variation for generalized nonlinear mixed-effects models that requires the specification of a null model is introduced as follows:

$$R_{VC}^2 = 1 - \frac{\sum_{i=1}^m (\mathbf{y}_i - \hat{\mathbf{y}}_i)'\boldsymbol{\Lambda}_i^{-1}(\mathbf{y}_i - \hat{\mathbf{y}}_i)}{\sum_{i=1}^m (\mathbf{y}_i - \hat{\mathbf{y}}_{i0})'\boldsymbol{\Lambda}_i^{-1}(\mathbf{y}_i - \hat{\mathbf{y}}_{i0})},$$

for any positive definite matrix $\boldsymbol{\Lambda}_i$. Sensible choices for $\boldsymbol{\Lambda}_i$ are either $\hat{\sigma}^2\mathbf{I}_{n_i}$ or $\hat{\sigma}_0^2\mathbf{I}_{n_i}$. To compare across different candidate models, the most reasonable choice is $\hat{\sigma}_0^2\mathbf{I}_{n_i}$.

This measure can be interpreted as the proportional decrease in residual variability compared with the residual variability of a null model, which is either one containing only a fixed intercept or one with a fixed intercept and a random intercept. As for r_c, a marginal and a conditional R_{VC}^2 can be defined, depending whether $\hat{\mathbf{b}}_i$ in $\hat{\mathbf{y}}_i$ are set to zero or not (original definition). The adjusted values are denoted $R_{VC,a}^2$ and are defined as $1 - N(N-p)^{-1}(1 - R_{VC}^2)$. As for r_c, the marginal R_{VC}^2 can be used for fixed effects selection and the adjusted conditional R_{VC}^2, $R_{VC,a}^2$, allows one for adopting it for model selection.

3.1.5. Zheng (2000)

For generalized LMMs with normally distributed random effects, we consider three measures for assessing model adequacy among a set proposed by Zheng (2000). Based on the notion of scaled deviance defined as $d(\mathbf{y}, \boldsymbol{\mu}) = -2\phi[l(\boldsymbol{\mu}, \phi; \mathbf{y}) - l(\mathbf{y}, \phi; \mathbf{y})]$, where $l(\boldsymbol{\mu}, \phi; \mathbf{y})$ is the log of the joint likelihood function given $\mathbf{X}$ and $\mathbf{b}$, for a fixed dispersion parameter ϕ, the $N \times 1$ vector of responses $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_m)'$, the $N \times p$ design matrix for fixed effects $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_m)'$, and the $mq \times 1$ vector of random effects $\mathbf{b} = (\mathbf{b}_1, \ldots, \mathbf{b}_m)'$, $\boldsymbol{\mu} = \text{E}(\mathbf{y} \mid \mathbf{X}, \mathbf{b})$, the first measure is the proportional reduction in deviance D_{rand}, defined as

$$D_{rand} = 1 - \frac{\sum_{i=1}^m d_i(\mathbf{y}_i, \hat{\mathbf{y}}_i)}{\sum_{i=1}^m d_i(\mathbf{y}_i, \bar{y}\mathbf{1}_{n_i})},$$

where the numerator is the deviance under the model of interest and the denominator is the deviance under the null model that fits only a fixed intercept.

D_{rand} is an extension of the coefficient of determination as, in the normal LMM, $\sum_{i=1}^{m} d_i(\mathbf{y}_i, \hat{\mathbf{y}}_i) = \sum_{i=1}^{m} (\mathbf{y}_i - \hat{\mathbf{y}}_i)'(\mathbf{y}_i - \hat{\mathbf{y}}_i)$ and $\sum_{i=1}^{m} d_i(\mathbf{y}_i, \bar{y}\mathbf{1}_{n_i}) = \sum_{i=1}^{m} (\mathbf{y}_i - \bar{y}\mathbf{1}_{n_i})'(\mathbf{y}_i - \bar{y}\mathbf{1}_{n_i})$. The measure D_{rand} ranges between 0 and 1, with 0 meaning that the model of interest provides no improvement in prediction over the null model and 1 meaning perfect prediction.

The second measure is the proportional reduction in Penalized Quasi-Likelihood (PQL) P_{rand}, defined as

$$P_{rand} = 1 - \frac{\sum_{i=1}^{m} d_i(\mathbf{y}_i, \hat{\mathbf{y}}_i)/(2\phi) + \hat{\mathbf{b}}'(\hat{\mathbf{D}} \otimes \mathbf{I}_m)^{-1}\hat{\mathbf{b}}/2}{\sum_{i=1}^{m} d_i(\mathbf{y}_i, \bar{y}\mathbf{1}_{n_i})/(2\phi)},$$

where $\hat{\mathbf{b}} = (\hat{\mathbf{b}}_1, \ldots, \hat{\mathbf{b}}_m)'$ is the $mq \times 1$ vector of estimated random effects. P_{rand} ranges between 0 and 1, equals 0 when the prediction is poor and/or the random effects are large, and equals 1 when the prediction is perfect and the random effects are 0. Due to the penalty for large random effects, P_{rand} allows for model selection.

The third measure by Zheng (2000) is the concordance index c, a measure of cross-sectional agreement. It equals the proportion of pairs of observations with unequal $\mathbf{y}$ values for which the ranks of $\hat{\mathbf{y}}$ and of $\mathbf{y}$ are concordant. As c depends on ranking information only, it cannot distinguish between models that yield the same ranking of the fitted values.

In order to identify the best set of fixed effects, we consider the marginal versions of these three measures obtained by setting $\hat{\mathbf{b}}_i = \mathbf{0}$ in $\hat{\mathbf{y}}_i$ for D_{rand} and c, and, for P_{rand}, by setting $\hat{\mathbf{b}}_i = \mathbf{0}$ in $\hat{\mathbf{y}}_i$ and in the second term of its numerator, which simplifies to the marginal D_{rand}, showing that P_{rand} is also an extension of the coefficient of determination.

3.1.6. Xu (2003)

Xu (2003) presented three measures that require a null model specified either with only a fixed intercept, or with also a random intercept. For both null models, $\sigma^2 = \sigma_0^2$.

The two measures of explained variation are defined as

$$r_X^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_0^2},$$

where $\hat{\sigma}^2$ and $\hat{\sigma}_0^2$ are the estimates of σ^2 and σ_0^2, and

$$R_X^2 = 1 - \frac{\sum_{i=1}^{m} (\mathbf{y}_i - \hat{\mathbf{y}}_i)'(\mathbf{y}_i - \hat{\mathbf{y}}_i)}{\sum_{i=1}^{m} (\mathbf{y}_i - \hat{\mathbf{y}}_{i0})'(\mathbf{y}_i - \hat{\mathbf{y}}_{i0})}.$$

A measure of explained randomness using the conditional likelihood of the observed data given the predicted random effects is also proposed as

$$\rho_X^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_0^2} \exp\left(\frac{\sum_{i=1}^{m} (\mathbf{y}_i - \hat{\mathbf{y}}_i)'(\mathbf{y}_i - \hat{\mathbf{y}}_i)}{N\hat{\sigma}^2} - \frac{\sum_{i=1}^{m} (\mathbf{y}_i - \hat{\mathbf{y}}_{i0})'(\mathbf{y}_i - \hat{\mathbf{y}}_{i0})}{N\hat{\sigma}_0^2} \right).$$

This measure of randomness is related with the likelihood ratio statistics $N\hat{\Gamma}$ for testing against the null model with intercept and random intercept only by the relationship $\rho_X^2 = 1 - \exp(-\hat{\Gamma})$.

All the three measures (with range [0,1]) are interpretable as the classical coefficient of determination. Xu (2003) emphasized that in ordinary linear regression, if REML estimates are used for the variance components, r_X^2 is the adjusted R^2. As an extension of the classical adjusted coefficient of determination, r_X^2 can be used for model selection.

For fixed effects selection, we consider the marginal version of R_X^2 by setting $\hat{\mathbf{b}}_i = \mathbf{0}$ in $\hat{\mathbf{y}}_i$.

3.1.7. Liu et al. (2008)

To estimate the portion of explained variation of the modeled data by the fitted LMM, Liu et al. (2008) introduced three R^2 statistics. The measures compare the residuals of an intercept-only model with the residuals of the model of interest considering, respectively, the fixed effects only (R_{F}^2), the fixed and random effects (R_{T}^2), and the total fixed effects where all variables and individuals are treated as fixed (R_{TF}^2):

$$R_{\mathrm{F}}^2 = 1 - \frac{\sum_{i=1}^{m} (\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})'(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})}{\sum_{i=1}^{m} (\mathbf{y}_i - \bar{y}\mathbf{1}_{n_i})'(\mathbf{y}_i - \bar{y}\mathbf{1}_{n_i})},$$

$$R_{\mathrm{T}}^2 = 1 - \frac{\sum_{i=1}^{m} (\mathbf{y}_i - \hat{\mathbf{y}}_i)'(\mathbf{y}_i - \hat{\mathbf{y}}_i)}{\sum_{i=1}^{m} (\mathbf{y}_i - \bar{y}\mathbf{1}_{n_i})'(\mathbf{y}_i - \bar{y}\mathbf{1}_{n_i})}$$

and

$$R_{\mathrm{TF}}^2 = 1 - \frac{\sum_{i=1}^{m} (\mathbf{y}_i - (\mathbf{X}_i, \mathbf{Z}_i)\hat{\boldsymbol{\eta}})'(\mathbf{y}_i - (\mathbf{X}_i, \mathbf{Z}_i)\hat{\boldsymbol{\eta}})}{\sum_{i=1}^{m} (\mathbf{y}_i - \bar{y}\mathbf{1}_{n_i})'(\mathbf{y}_i - \bar{y}\mathbf{1}_{n_i})},$$

where $\hat{\boldsymbol{\eta}} = \sum_{i=1}^{m} \left((\mathbf{X}_i, \mathbf{Z}_i)'(\mathbf{X}_i, \mathbf{Z}_i)\right)^{-1} (\mathbf{X}_i, \mathbf{Z}_i)'\mathbf{y}_i$. R_{F}^2 is obtained by setting $\hat{\mathbf{b}}_i = 0$ in R_{T}^2 and can be seen as a marginal measure to be used for fixed effects selection. Finally, R_{T}^2 can be seen as a conditional measure, as in Sections 3.1.3 and 3.1.4. Liu et al. (2008) noted that R_{TF}^2 is defined more as a theoretical measure of the upper bound of R^2 rather than a practical R^2 measure. We nevertheless consider it for completeness.

To adjust for the dimensions of the design matrices, adjusted values for R^2_{F} and R^2_{TF}, but not for R^2_{T}, are proposed: $R^2_{\mathrm{F},a} = 1 - N(N-(p+q))^{-1}(1-R^2_{\mathrm{F}})$ and $R^2_{\mathrm{TF},a} = 1 - N(N-\mathrm{rank}((\mathbf{X}_i,\mathbf{Z}_i)))^{-1}(1-R^2_{\mathrm{TF}})$. $R^2_{\mathrm{TF},a}$ can thus be used for model selection.

These measures usually range between 0, indicating complete lack of fit, and 1, indicating perfect fit. Negative values are also possible and would indicate that the model of interest provides no improvement, in terms of explained variation, over the intercept-only model.

3.1.8. Measures comparison

Because the y_{ij} are supposed normally distributed, some of the measures discussed above are equivalent. This is the case for the marginal R^2_{VC}, the marginal D_{rand}, the marginal P_{rand}, the marginal R^2_X and R^2_{F}; for the conditional R^2_{VC} and R^2_X based on the null model with a fixed intercept, D_{rand} and R^2_{T}; and for the conditional R^2_{VC} and R^2_X based on the null model with a fixed intercept and a random intercept. When we consider the adjusted values for the marginal R^2_{VC}, the conditional R^2_{VC}'s and R^2_{F}, these equalities no longer hold.

Some other measures, while they are not equivalent, are defined similarly and will thus give close values. This is the case for the two measures of Snijders and Bosker (1994); for the marginal r_c, the marginal R^2_{VC}, the marginal D_{rand}, the marginal P_{rand}, the marginal R^2_X and R^2_{F}; for the conditional r_c, the conditional R^2_{VC}, R^2_X and ρ^2_X based on the null model with a fixed intercept, D_{rand}, P_{rand}, R^2_{T} and R^2_{TF}; and for the conditional R^2_{VC}, R^2_X and ρ^2_X based on the null model with a fixed intercept and a random intercept.

3.2. Measures of model selection

We introduce in this Section the measures belonging to category S (cf. Table 1).

3.2.1. Akaike (1974) and Schwarz (1978) information criteria

The marginal AIC (mAIC, Akaike, 1974) is a general criterion for model selection defined by

$$\mathrm{mAIC} = -2\mathrm{LL} + 2k,$$

where k is the number of independently adjusted parameters and LL is the marginal log-likelihood. For ML estimation, k is the sum of the number of fixed effects p, the number of parameters of the covariance matrix $\mathbf{D}$ plus one because of the estimation of σ^2. For REML estimation, LL is replaced by the restricted log-likelihood and k is computed as the number of parameters of the random effects covariance matrix $\mathbf{D}$ plus one because of the estimation of σ^2 (e.g., Greven and Kneib, 2010). The mAIC is often returned by statistical software (e.g., `lmer()` in R and `proc mixed` in SAS), but `lmer()` in R does not report the mAIC based on the restricted log-likelihood as it cannot be used to compare models with different sets of fixed effects. With a focus on random effects selection, Greven and Kneib (2010) showed that the mAIC is a biased estimator of the Akaike information and favors smaller models without random effects; thus, they recommend not to use it. Sakamoto (2019) derived the expression of the asymptotic bias of the mAIC, which depends on the true variance structure, and proposed a simulation based procedure to estimate it. The mAIC has also been interpreted as an estimator of the squared prediction error (Efron, 2004; Säfken and Kneib, 2020).

For LMM, Vaida and Blanchard (2005) introduced the conditional AIC (cAIC) by replacing the marginal log-likelihood with the conditional log-likelihood, and k by a term related to the effective degrees of freedom proposed by Hodges and Sargent (2001). By relaxing the assumption of known covariance matrix of random effects, Liang et al. (2008) defined a generalized version of cAIC for LMMs considering a numerical approximation of the penalty function. To reduce the computational burden, Greven and Kneib (2010) derived an analytic representation of the generalized cAIC of Liang et al. (2008). Note that the cAIC can be computed under either ML and REML estimation, giving rise to different penalty terms.

As discussed in Vaida and Blanchard (2005), both the mAIC and the cAIC tend to favor complex models. By taking in consideration the total number of observations N, the BIC penalizes model complexity heavily and is defined by

$$\mathrm{BIC} = -2\mathrm{LL} + \log(N)k.$$

To our knowledge, a conditional version of the BIC does not exist, therefore BIC denotes here its marginal version.

Here, we consider the mAIC and the BIC for ML estimation, as it is widely used in practice due to its availability in standard statistical software, and the cAIC for ML and REML estimation (Greven and Kneib, 2010; Saefken et al., 2014).

The best model in terms of adjustment is the one with smallest AIC (or BIC) value and a difference between two AIC (or BIC) values is considered unimportant if less than 2 and important if greater or equal to 3-7 (see Burnham and Anderson, 2002).

3.2.2. Spiegelhalter et al. (2002)

The DIC is a Bayesian measure of fit to compare complex, possibly nonlinear, hierarchical models and is defined by

$$\mathrm{DIC} = D(\bar{\boldsymbol{\theta}}) + 2p_D, \tag{3}$$

where $\boldsymbol{\theta}$ are the unknown parameters, $\bar{\boldsymbol{\theta}} = \mathrm{E}(\boldsymbol{\theta} \mid \mathbf{y})$ is the posterior mean of $\boldsymbol{\theta}$ for observed data $\mathbf{y}$, and $p_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}) = \mathrm{E}_{\boldsymbol{\theta}|\mathbf{y}}(D(\boldsymbol{\theta})) - D(\bar{\boldsymbol{\theta}})$ is the effective number of parameters. $D(\boldsymbol{\theta})$ is the Bayesian deviance and is defined as $-2\log(p(\mathbf{y} \mid \boldsymbol{\theta})) + 2\log(f(\mathbf{y}))$, where $p(\mathbf{y} \mid \boldsymbol{\theta})$ is the probability model and $f(\mathbf{y})$ is some standardizing term that is a function of the data

alone. $D(\boldsymbol{\theta})$ is the saturated deviance when $f(\mathbf{y}) = p(\mathbf{y} \mid \mu(\boldsymbol{\theta}) = \mathbf{y})$, for members of the exponential family with $\mathrm{E}(\mathbf{Y}) = \mu(\boldsymbol{\theta})$, where $\mathbf{Y}$ are unobserved future data introduced by the authors to posit a "true" distribution $p^t(\mathbf{Y})$.

When comparing alternative models, the best in terms of adjustment has the smallest DIC and the rules of thumb to claim for important differences in DIC are the same as for the AIC (cf. Section 3.2.1).

3.3. Implementation

We have done all of our computation in R (R Core Team, 2020). All the scripts are available on the GitHub site https://github.com/ecantoni/R2_LMM.

To compute the measures presented in the frequentist framework, models are fitted with the function `lmer()` from the lme4 package (Bates et al., 2015), to obtain either REML or ML estimates. The measures of R^2 are easily computed from the fitted models and we provide our own implementation of their computation. At the best of our knowledge, only few of them are available in R packages. The measure of (Snijders and Bosker, 1994) were available in the past in the `r2` function of package sjstats, but this function is now deprecated. The package performance contains a generic `r2` function that defaults to the measure of (Nakagawa et al., 2017) for linear mixed models. In the same package, a function `r2_xu` computes the measure of (Xu, 2003). Some of the measures of (Liu et al., 2008) are available in function `rsq` of package rsq. The marginal AIC and BIC are obtained by the built-in functions `AIC` (in case of REML estimation) and `extractAIC` (in case of ML estimation). The conditional AIC is implemented in the R package cAIC4 (Saefken et al., 2014).

In the Bayesian framework, the function `stan()` from the package rstan (Stan Development Team, 2018) has been used to fit the models. We provide our own code to compute the DIC and the (Gelman and Pardoe, 2006) measures.

4. Home radon levels

4.1. Description of data and fitted models

The home radon levels data from (Gelman and Pardoe, 2006) consist in measurements of levels of radon gas in $N = 919$ houses clustered within $m = 85$ counties in Minnesota, USA. Radon is a radioactive gas that forms naturally in the soil by decay of uranium. When breathing, radon can enter the lungs, settle on the lung tissue and irradiate it, resulting possibly in a lung cancer. The distribution of radon levels varies greatly from one house to another and these measurements aim at identifying areas with high radon exposures. The number of measures n_i per county varies from 1 to 116. Two predictors are used, a house predictor at the first level of the hierarchical model and a county predictor at the second level. The former indicates whether the measurement was taken in a basement (basement_{ij}=1) or on the first floor (basement_{ij}=0). The latter measures the soil uranium content in each county (uranium_i). The logarithm of the radon measurement (y_{ij}) is modeled in the analyses.

Seven models are fitted to these highly unbalanced data for which we compute all measures presented in Section 3. The most complex model is the varying-intercept and varying-slope model of (Gelman and Pardoe, 2006) defined by

$$\begin{aligned} y_{ij} &\sim \mathcal{N} \quad (\alpha_i + \beta_i \text{basement}_{ij}, \sigma^2), \\ \begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} &\sim \mathcal{N} \quad \left(\begin{pmatrix} \gamma_0 + \gamma_1 \text{uranium}_i \\ \delta_0 + \delta_1 \text{uranium}_i \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix} \right), \end{aligned} \tag{4}$$

for $i = 1, \ldots, 85$ and $j = 1, \ldots, n_i$. We call this model 6. Its fixed part contains the predictors basement, uranium (predictor for the random intercept) and the cross-level interaction between basement and uranium (predictor for the random slope), whereas its random part contains a random intercept and a random slope for basement. We consider the following alternative models, all nested within model 6 :

- Model 0 contains only a fixed intercept ($\alpha_i = \alpha \ \forall i$ and $\beta_i = 0 \ \forall i$).
- Model 1 contains no random effect and contains neither the uranium predictor nor the cross-level interaction and is thus a simple linear regression model with the predictor basement ($\alpha_i = \alpha \ \forall i$ and $\beta_i = \beta \ \forall i$).
- Model 2 contains only a fixed intercept and a random intercept ($\gamma_1 = 0$ and $\beta_i = 0 \ \forall i$).
- Model 3 corresponds to model 6 without the uranium predictor and the cross-level interaction ($\gamma_1 = \delta_1 = 0$).
- Model 4 corresponds to model 6 without the cross-level interaction ($\delta_1 = 0$).
- Model 5 corresponds to model 6 with non-correlated random effects ($\rho = 0$).

We aim at evaluating the overall adequacy of these different models considering the values of the measures belonging to categories A and A&S. Furthermore, we would like to identify the best set of fixed effects with the measures in category F&S and identify the model that fits best these data using the measures of model selection (category S). We expect the measures in a same category to give similar conclusions. For the model that fits best the data, we will give insights of what influences the levels of radon gas.

4.2. Results

Table 2 presents the values of the measures of interest for all the models considered. To facilitate the reading of the Table we boldfaced the "'best" value for each measure, but we warn the reader that in many cases there is no clear winner

Table 2
Measures for models of home radon levels data. For the measures computed with several null models, (0) indicates that the considered null model is model 0 and (2) indicates that the considered null model is model 2. The dashed line separates null models. m.=marginal; c.=conditional. "Best" value for each measure highlighted in bold.

			Model						
Reference [Section]	Measure	Category	0	1	2	3	4	5	6
Gelman and Pardoe (2006) [3.1.1]	R^2_y	A		0.071	0.125	**0.228**	0.226	0.214	0.222
	λ_y	A		0.001	0.044	**0.063**	0.041	0.029	0.031
	R^2_α	A			0	0	0.397	**0.517**	0.239
	λ_α	A			0.534	0.647	0.782	**0.818**	0.725
	R^2_β	A				0	0	**0.757**	0.377
	λ_β	A				0.870	0.861	**0.942**	0.836
Zheng (2000) [3.1.5]	D_{rand}	A		0.072	0.163	**0.285**	0.260	0.232	0.256
	c	A		0.567	0.653	**0.690**	0.682	0.675	0.679
Xu (2003) [3.1.6]	R^2_X (0)	A		0.072	0.163	**0.285**	0.260	0.232	0.256
	ρ^2_X (0)	A		0.072	0.162	**0.283**	0.260	0.231	0.255
	R^2_X (2)	A				**0.145**	0.116	0.082	0.111
	ρ^2_X (2)	A				**0.144**	0.116	0.082	0.111
Liu et al. (2008) [3.1.7]	R^2_{T}	A		0.072	0.163	**0.285**	0.260	0.232	0.256
Vonesh et al. (1996) [3.1.3]	c. $r_{c,a}$	A&S		0.132	0.230	**0.397**	0.384	0.356	0.380
Vonesh and Chinchilli (1996) [3.1.4]	c. $R^2_{VC,a}$ (0)	A&S		0.070	0.162	**0.283**	0.258	0.228	0.253
	c. $R^2_{VC,a}$ (2)	A&S				**0.143**	0.113	0.078	0.107
Zheng (2000) [3.1.5]	P_{rand}	A&S		0.072	0.126	**0.236**	0.232	0.215	0.231
Xu (2003) [3.1.6]	r^2_X (0)	A&S		0.071	0.126	**0.235**	0.231	0.213	0.229
	r^2_X (2)	A&S				**0.125**	0.120	0.100	0.118
Liu et al. (2008) [3.1.7]	$R^2_{\mathrm{TF},a}$	A&S		0.070	0.124	NA	**0.233**	**0.233**	**0.233**
Snijders and Bosker (1994) [3.1.2]	R^2_1	F&S				0.026	0.136	**0.185**	0.152
	R^2_2	F&S				-0.148	0.231	**0.379**	0.272
Vonesh et al. (1996) [3.1.3]	m. $r_{c,a}$	F&S		0.132	-0.001	0.142	0.304	**0.307**	0.303
Vonesh and Chinchilli (1996) [3.1.4]	m. $R^2_{VC,a}$	F&S		0.070	-0.012	0.047	0.179	**0.182**	**0.182**
Zheng (2000) [3.1.5]	m. D_{rand}	F&S		0.072	-0.011	0.049	0.182	**0.185**	**0.185**
	m. P_{rand}	F&S		0.072	-0.011	0.049	0.182	**0.185**	**0.185**
	m. c	F&S		0.554	0.500	0.554	**0.656**	0.655	0.655
Xu (2003) [3.1.6]	m. R^2_X	F&S		0.072	-0.011	0.049	0.182	**0.185**	**0.185**
Liu et al. (2008) [3.1.7]	$R^2_{\mathrm{F},a}$	F&S		0.070	-0.013	0.045	0.177	**0.180**	**0.180**
	-2LL ML	S	2315.479	2247.025	2255.237	2161.109	2117.603	2118.030	**2114.224**
	-2LL REML	S			2259.442	2168.325	2128.640	2130.906	**2126.579**
Akaike (1974) [3.2.1]	mAIC	S	2319.479	2253.025	2261.237	2173.109	2131.603	2132.030	**2130.224**
Schwarz (1978) [3.2.1]	BIC	S	2329.126	2267.495	2275.707	2202.048	**2165.366**	2165.793	2168.810
Vaida and Blanchard (2005) [3.2.1]	cAIC ML	S			2237.164	2139.145	**2121.134**	2124.228	2122.753
	cAIC REML	S			2237.141	2138.719	**2120.713**	2123.962	2121.438
Spiegelhalter et al. (2002) [3.2.2]	DIC	S	2319.409	2253.077	2236.730	2139.035	2120.149	2124.989	**2115.792**

(i.e., differences in measure values across models are often minor and not fully interpretable). The ML estimation is used for -2LL ML, the mAIC, the BIC and the cAIC ML, and the REML estimation is used for all the others measures. For the measures presented in the Bayesian framework, the models are estimated with 40000 iterations and 3 chains.

To evaluate overall model adequacy, we consider the measures belonging to categories A and A&S. The obtained values are really close for R^2_y, the conditional $R^2_{\mathrm{VC},a}$ based on the null model 0, D_{rand}, P_{rand}, r^2_X, R^2_X and ρ^2_X based on the null model 0, R^2_{T} and $R^2_{\mathrm{TF},a}$, that are all extensions of the classical R^2. These measures indicate that models 3, 4, 5 and 6 explain around 20% (or slightly more) of the variation in the data. $R^2_{\mathrm{TF},a}$ is equal for models 4, 5 and 6, and for model 3, this measure cannot be computed because $\left((\mathbf{X}_i, \mathbf{Z}_i)'(\mathbf{X}_i, \mathbf{Z}_i)\right)$ is singular (as noted in Section 3.1.7, this measure is more a theoretical measure).

Table 3
Estimated coefficients (by REML) for model 4 (given by Equation (4) with $\delta_1 = 0$) for the radon dataset.

γ_0	γ_1	δ_0	ρ	σ_α	σ_β	σ
0.820	0.642	0.768	-0.950	0.404	0.357	0.748

In comparison with the measures listed above, which include R^2_y, the other measures of (Gelman and Pardoe, 2006) give additional information and have thus different values. Indeed, for the first "level," the pooling factor λ_y moreover indicates that the within-group sample sizes are reasonably large as λ_y is close to zero for all models. Furthermore, for the second "level," R^2_α gives the percentage of the variation among counties that is explained by the uranium level (for instance, there is around 40% in model 4), and λ_α informs that there is higher population-level information than within-county information, as λ_α is higher than 50%. For the third "level," R^2_β gives the percentage of the variation in the basement effects across counties that is explained by the uranium level (for instance, there is around 38% in model 6). And finally, λ_β being close to 90% for all the models containing a random slope means that the individual counties do not add information compared to the county-level model. As c and the conditional $r_{c,a}$ are interpreted as concordance coefficients, their values are different (in this example, they are higher) from the values of the extensions of the classical R^2.

To summarize, even if some values are different, all these measures in categories A and A&S, except $R^2_{\text{TF},a}$, indicate that models 3, 4 and 6 provide a similar fit to the data, whatever the considered null model. Indeed, the difference is that the values of the measures with the null model 2 are smaller than with the null model 0, as the proportion of variability in the model of interest compared with the variability of the null model 2 is smaller than the same proportion for the null model 0. For these measures of overall model adequacy, model 5, assuming non-correlated random effects, does not fit the data well. This seems reasonable in view of the large correlation estimate of -0.950 in both model 4 and 6.

The measures allowing evaluating model adequacy due to fixed effects and fixed effects selection (category F&S) are sometimes negative, indicating that the fixed part of the corresponding model is misspecified. As for the measures of overall model adequacy, there are some differences in the obtained values but the choice of the best set of fixed effects is not altered by these differences. By comparing the values for the different models, we conclude that the predictors basement and uranium are useful but the cross-level interaction between basement and uranium is superfluous. Indeed the addition of this predictor in models 5 and 6 has no influence on the value of the measures.

The values of the measures of model selection (category S) are close, as expected. According to the measures from the category F&S, we can conclude that the cross-level interaction does not contribute to the adjustment. Based on information criteria (mAIC, BIC, cAIC ML, cAIC REML and DIC), models 4, 5 and 6 cannot easily be distinguished even if the number of parameters is penalized. Thus, based on these observations and on parsimony, we choose model 4 as the most appropriate.

To conclude, the measures of overall model adequacy give similar results and indicate models 3, 4 and 6 as the most adequate for these data. Despite similar results, the interpretation of these measures (categories A and A&S), and of those allowing evaluating model adequacy due to fixed effects (category F&S), can be different (concordance coefficient vs. explained variation). Thus, researchers should be wary when using such measures. Given the values of the measures allowing model selection (categories F&S and S), model 4 (defined by Equation (4) with $\delta_1 = 0$) seems the best for these data. Table 3 gives the estimated coefficients of this model in the frequentist approach with REML. It thus seems that levels of radon gas are higher for both houses with basement and counties with higher soil uranium content. Moreover, the levels of radon gas are different among counties and the effect of having a house with a basement on the level of radon gas varies among counties.

5. Simulation study

We conduct a simulation study in order to compare, within the different categories given in Table 1, the considered measures. In particular, we first test the sensitivity of all the measures of interest to modifications of model parameters. Indeed, we expect the modification of model parameters to impact the values of the measures. For instance, if the modification of a model parameter increases the variability in the simulated data, a R^2-type measure should be larger as there is more variability to explain. Second, we test the ability of the measures designed for model selection (categories A&S, F&S and S) to identify the correct set of fixed effects or the correct model, respectively, among a series of seven alternatives. Below, we present the design of the simulation and the results.

5.1. Simulation study design

We considered 32 different simulation cases based on a full 2^5 factorial design. Within each case, we generated 200 samples from model 6 defined in (4). We thus used the covariates values from the home radon levels data and simulated normally distributed random effects and errors to generate outcomes. We estimated the seven models presented in Section 4 and computed the different measures from Section 3. As the model from which we generated the data is the largest (model 6), we do not evaluate the ability of the considered measures to select larger models.

The full simulation design is presented in Table 4. The five parameters ρ, γ_1 (associated to the predictor uranium), δ_1 (associated to the cross-level interaction between basement and uranium), σ_α and σ_β take each two different values. To

Table 4
Simulation cases. The original values of the parameters are those of case 1, in bold. Multiple empty rows refer to the value indicated between their delimiting horizontal lines (e.g., for Cases 1-8, $\sigma_\alpha = 0.268$)

	Parameter's value				
Case	ρ	σ_α	γ_1	σ_β	δ_1
1					**0.410**
2				**0.197**	0.205
3			**0.391**		0.410
4				0.230	0.205
5		**0.268**			0.410
6				0.197	0.205
7			0.195		0.410
8				0.230	0.205
9	**-0.820**				0.410
10				0.197	0.205
11			0.391		0.410
12				0.230	0.205
13		0.310			0.410
14				0.197	0.205
15			0.195		0.410
16				0.230	0.205
17					0.410
18				0.197	0.205
19			0.391		0.410
20				0.230	0.205
21		0.268			0.410
22				0.197	0.205
23			0.195		0.410
24				0.230	0.205
25	0				0.410
26				0.197	0.205
27			0.391		0.410
28				0.230	0.205
29		0.310			0.410
30				0.195	0.205
31			0.197		0.410
32				0.230	0.205

fix the values of the parameters, we estimated in the Bayesian framework 200 times models 3 and 6 on the original data. The values for the parameters γ_0, δ_0 and σ are set at 0.778, 0.692 and 0.753, corresponding to the median of the 200 point estimates obtained from the estimations of model 6, and are kept unchanged. The first value of the other parameters is fixed to the median of the 200 point estimates obtained from the estimations of model 6, which we call the original value ($\rho = -0.820$, $\gamma_1 = 0.391$, $\delta_1 = 0.410$, $\sigma_\alpha = 0.268$ and $\sigma_\beta = 0.197$). The second value for ρ is 0, representing the case of non-correlated random effects. The modified value for γ_1 is fixed to the original value divided by two (0.195), yielding a

Table 5
Spearman correlations matrix for measures in categories A and A&S computed after estimation of model 6 in simulation case 1 (original values). For the measures computed with several null models, (0) indicates that the considered null model is model 0 and (2) indicates that the considered null model is model 2. c.=conditional.

	c. $r_{c,a}$	c. $R^2_{VC,a}(0)$	D_{rand}	P_{rand}	c	$r^2_X(0)$	$R^2_X(0)$	$\rho^2_X(0)$	R^2_T	$R^2_{TF,a}$	c. $R^2_{VC,a}(2)$	$r^2_X(2)$	$R^2_X(2)$	$\rho^2_X(2)$
c. $r_{c,a}$	1	0.990	0.990	0.998	0.947	0.998	0.990	0.990	0.990	0.941	0.445	0.508	0.445	0.449
c. $R^2_{VC,a}$		1	1	0.981	0.945	0.981	1	1	1	0.933	0.477	0.510	0.477	0.480
D_{rand}			1	0.981	0.945	0.981	1	1	1	0.933	0.477	0.510	0.477	0.480
P_{rand}				1	0.945	1	0.981	0.982	0.981	0.940	0.429	0.502	0.429	0.433
c					1	0.945	0.945	0.946	0.945	0.892	0.340	0.380	0.340	0.343
r^2_X						1	0.981	0.982	0.981	0.940	0.429	0.502	0.429	0.433
R^2_X							1	1	1	0.933	0.477	0.510	0.477	0.480
ρ^2_X								1	1	0.934	0.475	0.509	0.475	0.478
R^2_T									1	0.933	0.477	0.510	0.477	0.480
$R^2_{TF,a}$										1	0.354	0.417	0.354	0.357
c. $R^2_{VC,a}(2)$											1	0.962	1	1
$r^2_X(2)$												1	0.962	0.964
$R^2_X(2)$													1	1
$\rho^2_X(2)$														1

smaller signal-to-noise ratio for the intercept variation (7.3% vs. 24% of intercept variation explained by uranium). Similarly, the alternative δ_1 is fixed to the original value divided by two (0.205), reducing the signal-to-noise ratio for the slope from 39% to 13.8%. For σ_α, we choose an alternative value of 0.310, the median between the original value and the median of the 200 point estimates obtained from the estimations of model 3. Likewise, the alternative value for σ_β is 0.230.

5.2. Results

In Section 5.2.1, we present the correlations between the frequentist measures. In the subsequent Sections, we first compare the measures of interest within their category in terms of sensitivity to modifications of model parameters. To do so, we computed a separate full analysis of variance (ANOVA) for each measure for model 6 across all 32 simulation cases. The dependent variable of the ANOVAs is, in turn, each measure and the factors are the five parameters ρ, γ_1, δ_1, σ_α, and σ_β, and their interactions. Each factor has two levels; 0 for the original value and 1 for the modified value. Remark that the variability in the simulated data is higher when γ_1 and δ_1 are equal to their original values and when ρ, σ_α and σ_β are equal to their modified values. This means that the variability is higher for bigger values of γ_1, δ_1, σ_α and σ_β and for $\rho = 0$. All the results are summarized in Table 8 in terms of the effect size partial η^2 (Cohen, 1988) ($SS_{effect}/(SS_{effect} + SS_{error\ for\ that\ effect}$), where SS_{effect} are the sums of squares of each effect and $SS_{error\ for\ that\ effect}$ those of the error for that effect). Table 8 includes only significant terms, which were all main effects and some two-way interactions (i.e., all higher order interactions were non significant). Considering the adjusted or the non adjusted measure as dependent variable will give the same ANOVA results. For the measures of model selection (categories A&S, F&S and S), we evaluate them in terms of performance in model selection using a systematic rule to identify the selected model based on the best value of each measure (either biggest or smallest, depending on the measure). Table 9 shows how often each measure chooses one of the competing models for simulation cases 1 and 2. The results presented in Tables 8 and 9 are discussed by category of measures in the corresponding Section.

5.2.1. Correlations

For the frequentist measures, we computed Spearman correlations by category. Table 5 presents the correlations for the measures in categories A and A&S. Pearson correlations (not shown) are virtually identical.

The measures based on the same null model are highly related but the same measures based on a different null model are not correlated. Indeed, the measures based on the null model 2 assess the percentage of variation explained by covariates conditional on the cluster randomization design, while the measures based on the null model 0 assess the percentage of variation explained by covariates and clustering (Xu, 2003). Table 6 gives the correlations for the measures belonging to the category F&S. These measures are highly related except R^2_2 that correlates less than the others, probably because R^2_2 identifies which predictors are useful to predict $\bar{y}_{i.}$ instead of y_{ij}. Finally, the correlations between the measures of the category S, that are all close to 1, are given in Table 7. These preliminary results confirm the expectations based on the measures definitions (cf. Section 3.1.8) and on the measures categories (cf. Table 1).

5.2.2. Comparison of measures of overall model adequacy

The first stage of the analysis of the results is to compare the measures allowing for evaluating overall model adequacy (categories A and A&S) in terms of sensitivity to modifications of model parameters. This comparison is done separately for λ_y, R^2_α, λ_α, R^2_β and λ_β because they summarize information for each variance component, instead of giving one value for the whole model. And this is also done separately for the measures based on the null model 2 because they measure a different

Table 6
Spearman correlations matrix for the measures in category F&S computed after estimation of model 6 in simulation case 1 (original values). m.=marginal.

	R^2_1	R^2_2	m. $r_{c,a}$	m. $R^2_{VC,a}$	m. D_{rand}	m. P_{rand}	m. c	m. R^2_X	$R^2_{F,a}$
R^2_1	1	0.895	0.765	0.711	0.711	0.711	0.684	0.711	0.711
R^2_2		1	0.562	0.522	0.522	0.522	0.527	0.522	0.522
m. $r_{c,a}$			1	0.973	0.973	0.973	0.937	0.973	0.973
m. $R^2_{VC,a}$				1	1	1	0.957	1	1
m. D_{rand}					1	1	0.957	1	1
m. P_{rand}						1	0.957	1	1
m. c							1	0.957	0.957
m. R^2_X								1	1
$R^2_{F,a}$									1

Table 7
Spearman correlations matrix for the measures in category S computed after estimation of model 6 in simulation case 1 (original values).

	-2LL ML	-2LL REML	mAIC	BIC	cAIC ML	cAIC REML
-2LL ML	1	1	1	1	0.969	0.984
-2LL REML		1	1	1	0.969	0.985
mAIC			1	1	0.969	0.984
BIC				1	0.969	0.984
cAIC ML					1	0.969
cAIC REML						1

quantity from those based on the null model 0 (cf. Section 5.2.1). The comments of the three subsequent paragraphs refer to the measures categorized as A and A&S in Table 8. The other measures are discussed in Sections 5.2.3 and 5.2.4.

The measures of overall model adequacy (except λ_y, R^2_α, λ_α, R^2_β and λ_β and those based on the null model 2) have similar sensitivity to the modifications of the parameters. Indeed, these measures are sensitive to modifications of the parameters' values. We observe that the factor ρ has a sizeable effect (partial $\eta^2 \simeq 0.5$) and the two-way interactions involving ρ are also often significant. Moreover, the two-way interaction $\gamma_1 \times \delta_1$ is also often significant, as γ_1 and δ_1 often have a moderate effect (partial η^2 close to 0.1 for γ_1 and to 0.07 for δ_1). Conversely to ρ, the factor σ_α has a slight effect (partial $\eta^2 \simeq 0.04$) and σ_β even has a smaller effect than the factor σ_α (partial $\eta^2 \simeq 0.003$) on all the measures. To more deeply understand the results, we plotted the boxplots obtained in the ANOVAs. As they are similar for the measures allowing for assessing the overall adequacy of the model (except λ_y, R^2_α, λ_α, R^2_β and λ_β and those based on the null model 2), we present them only for R^2_y (cf. Figure 1), which is representative. When γ_1, δ_1 and σ_α are bigger, and when $\rho = 0$, there is more variability to explain and the values of R^2_y are bigger, as expected. When σ_β is bigger, R^2_y is bigger only when $\rho = 0$. Indeed when ρ is nonzero, σ_α and σ_β are linked and most of the variability remains in σ_α, while when $\rho = 0$, the variability is independently divided between σ_α and σ_β.

The other measures of (Gelman and Pardoe, 2006), λ_y, R^2_α, λ_α, R^2_β and λ_β, should be sensitive to the parameters at their corresponding "level." Moreover, $\lambda_{\text{"level"}}$ is computed using the information contained in the errors and should thus not be sensitive to the parameters of the fixed part of the model. As expected, the pooling factor λ_y is sensitive to the modifications of ρ, σ_α and σ_β. However, regardless of the value of these parameters, λ_y is always close to zero as it should always be the case (Gelman and Pardoe, 2006). As expected, R^2_α is sensitive to the different values of ρ, γ_1 and σ_α. We expected the pooling factor λ_α to be sensitive to the modifications on ρ and σ_α but it is also sensitive to σ_β, because of the strong correlation between random intercept and random slope. Indeed, when $\rho = 0$, the values of λ_α remain the same, independently of σ_β (cf. supplementary Figure 1). When γ_1 is bigger, the predictor uranium explains more of the variability between counties and thus R^2_α is bigger (cf. supplementary Figure 2). When σ_α is bigger, it increases the variability at that "level," and thus R^2_α is smaller and λ_α is closer to zero, because the estimated random effects are further from the mean. As explained for R^2_y, when $\rho = -0.820$, σ_α and σ_β are strongly associated and most of the variability remains in σ_α. Thereby, there is more variability to explain and R^2_α is bigger. This leads the estimated random effects to be closer to the mean and λ_α closer to one. For R^2_β and the pooling factor λ_β, the significant effects are ρ, δ_1 and σ_β. For λ_β, δ_1 is significant, which is surprising. Nevertheless, the effect size is tiny (partial $\eta^2 < 0.001$) and this difference is superficial (cf. supplementary Figure 3). When δ_1 is bigger, the predictor for the random slope, which corresponds to the cross-level interaction between basement and uranium, explains more of the variability in the basement effects across counties and R^2_β is thus bigger (cf. supplementary Figure 4). Increasing the value of σ_β increases the variability at that "level," resulting in a smaller R^2_β and a pooling factor λ_β closer to zero, as the estimated random effects are further from the mean value. More information is available when ρ is nonzero, in which case R^2_β is bigger and λ_β is closer to 1.

Table 8
Results of the analyses of variance: the factors are in the columns and the dependent variables are in the lines. The cells contain the partial η^2 effect size measure, partial η^2 smaller than 0.001 are indicated by the $\times$ symbol and non significant results (5% threshold) are in parentheses. For the measures computed with several null models, (0) indicates that the considered null model is model 0 and (2) indicates that the considered null model is model 2. The dashed line separates null models. m.=marginal; c.=conditional.

Measure	Category	ρ	γ_1	δ_1	σ_α	σ_β	$\rho \times \gamma_1$	$\rho \times \delta_1$	$\gamma_1 \times \delta_1$	$\rho \times \sigma_\alpha$	$\rho \times \sigma_\beta$
R^2_y	A	0.496	0.113	0.068	0.039	0.002	0.003	0.002	0.002	0.003	0.005
λ_y	A	0.639	(×)	(×)	0.085	0.002	(×)	(×)	(×)	(×)	0.005
R^2_α	A	0.046	0.117	(×)	0.028	(×)	×	×	(×)	(×)	(×)
λ_α	A	0.344	(×)	(×)	0.095	0.004	(×)	×	(×)	0.002	0.001
R^2_β	A	0.045	(×)	0.127	(×)	0.005	(×)	(×)	(×)	0.001	(×)
λ_β	A	0.220	(×)	×	(×)	0.016	(×)	(×)	(×)	0.003	0.004
D_{rand}	A	0.561	0.088	0.063	0.050	0.003	0.002	0.001	0.002	×	0.007
c	A	0.598	0.095	0.080	0.043	0.002	0.011	0.010	0.002	(×)	0.006
$R^2_X(0)$	A	0.561	0.088	0.063	0.050	0.003	0.002	0.001	0.002	×	0.007
$\rho^2_X(0)$	A	0.559	0.089	0.064	0.050	0.003	0.002	0.001	0.002	0.001	0.007
$R^2_X(2)$	A	0.008	0.012	0.016	(×)	0.004	0.002	×	(×)	(×)	(×)
$\rho^2_X(2)$	A	0.007	0.015	0.015	(×)	0.004	0.002	×	(×)	(×)	(×)
R^2_{T}	A	0.561	0.088	0.063	0.050	0.003	0.002	0.001	0.002	×	0.007
c. $r_{c,a}$	A&S	0.532	0.106	0.076	0.044	0.002	0.005	0.003	0.002	×	0.006
c. $R^2_{VC,a}(0)$	A&S	0.561	0.088	0.063	0.052	0.003	0.002	0.001	0.002	×	0.007
c. $R^2_{VC,a}(2)$	A&S	0.008	0.012	0.016	(×)	0.004	0.002	×	(×)	(×)	(×)
P_{rand}	A&S	0.508	0.104	0.075	0.041	0.003	0.002	0.001	0.002	0.002	0.006
r^2 (0)	A&S	0.508	0.104	0.074	0.041	0.003	0.002	0.001	0.002	0.002	0.006
r^2 (2)	A&S	0.003	(×)	0.002	(×)	0.003	(×)	(×)	(×)	(×)	(×)
$R^2_{\mathrm{TF},a}$	A&S	0.498	0.099	0.071	0.037	0.003	0.002	0.001	0.002	0.002	0.005
R^2_1	F&S	0.054	0.207	0.139	(×)	0.006	0.001	×	0.003	(×)	(×)
R^2_2	F&S	0.099	0.241	0.222	0.001	0.007	0.006	0.004	×	(×)	(×)
m. $r_{c,a}$	F&S	0.096	0.266	0.192	0.003	(×)	×	×	0.005	(×)	(×)
m. $R^2_{VC,a}$	F&S	0.097	0.205	0.145	0.003	(×)	0.001	×	0.005	(×)	(×)
m. D_{rand}	F&S	0.097	0.205	0.145	0.003	(×)	0.001	×	0.005	(×)	(×)
m. P_{rand}	F&S	0.097	0.205	0.145	0.003	(×)	0.001	×	0.005	(×)	(×)
m. c	F&S	0.056	0.256	0.229	0.002	(×)	(×)	(×)	(×)	(×)	(×)
m. R^2_X	F&S	0.097	0.205	0.145	0.003	(×)	0.001	×	0.005	(×)	(×)
$R^2_{\mathrm{F},a}$	F&S	0.097	0.205	0.145	0.003	(×)	0.001	×	0.005	(×)	(×)
-2LL ML	S	0.168	(×)	(×)	0.009	×	(×)	(×)	(×)	(×)	×
-2LL REML	S	0.160	(×)	(×)	0.009	×	(×)	(×)	(×)	(×)	×
mAIC	S	0.168	(×)	(×)	0.009	×	(×)	(×)	(×)	(×)	×
BIC	S	0.168	(×)	(×)	0.009	×	(×)	(×)	(×)	(×)	×
cAIC ML	S	0.050	(×)	(×)	0.002	(×)	(×)	(×)	(×)	(×)	(×)
cAIC REML	S	0.049	(×)	(×)	0.002	(×)	(×)	(×)	(×)	(×)	(×)
DIC	S	0.184	0.011	0.010	0.007	(×)	(×)	(×)	0.002	0.003	0.001

Concerning the measures based on the null model 2, we want to highlight that their behavior is similar, but with smaller effects, to their counterparts based on the null model 0 when we manipulate ρ and σ_β. However, the null model 2 contains a random intercept, thus the measures are not sensitive to the modifications of the values of σ_α. We observe also that the parameters γ_1 and δ_1 have a slight effect (partial η^2 close to 0.01) and bigger values of these parameters give rise to smaller values of the measures (cf. supplementary Figure 5). A different sensitivity to the modifications of the parameters is observed for r^2_X, for which γ_1 is not significant and δ_1 has a smaller effect size (partial $\eta^2 = 0.002$). It thus seems that r^2_X based on the null model 2 is less sensitive to the modifications of the parameters associated with the fixed part of the

Table 9
Number of times that the measure chooses model o ($o = 1, 2, 3, 4, 5, 6$) in simulation cases 1 and 2. m.=marginal; c.=conditional.

		Simulation case 1						Simulation case 2					
		Model						Model					
Measure	Category	1	2	3	4	5	6	1	2	3	4	5	6
c. $r_{c,a}$	A&S			139	22	4	35			127	24	4	45
c. $R^2_{VC,a}$	A&S			199	1					188	3		9
P_{rand}	A&S			69	19	10	102			68	15	10	107
r^2_X	A&S			89	44	6	61			90	51	5	54
R^2_1	F&S				9	178	13				18	174	8
R^2_2	F&S				14	175	10				15	176	9
m. $r_{c,a}$	F&S				40	91	69				69	86	45
m. $R^2_{VC,a}$	F&S				33	94	73				94	60	46
m. D_{rand}	F&S				9	107	84				20	103	77
m. P_{rand}	F&S				9	107	84				20	103	77
m. c	F&S				52	76	72				107	65	28
m. R^2_X	F&S				9	107	84				20	103	77
$R^2_{F,a}$	F&S				33	94	73				94	60	46
-2LL ML	S						200						200
-2LL REML	S				41		159				118		82
mAIC	S				48	99	53				108	73	19
BIC	S				76	122	2	1			121	78	
cAIC ML	S			1	78	48	73			3	120	34	43
cAIC REML	S			1	64	55	80			1	116	43	40
DIC	S		1	193	3		3			199			1

model. The boxplots of the measures based on the null model 2 are virtually identical to the supplementary Figure 5 and are thus not shown.

The second stage of the analysis of the results is to evaluate the measures of overall adequacy further allowing model selection (category A&S) in terms of performance in model selection. Here we comment the measures categorized as A&S in Table 9. The other measures are discussed in Sections 5.2.3 and 5.2.4. The reasons why these measures can further be used for model selection are that the conditional $r_{c,a}$ and $R^2_{VC,a}$ are adjusted for the number of parameters, P_{rand} penalizes for large random effects, and r^2_X is an extension of the classical adjusted R^2 (cf. Section 3.1.6). As seen in Section 5.2, $R^2_{\text{TF},a}$ cannot be computed for model 3 and is equal for models 4, 5 and 6. Moreover, this measure is conceived as a theoretical measure of the upper bound of R^2_X (cf. Section 3.1.7). Thereby, we will not use $R^2_{\text{TF},a}$ for model selection. Considering the remaining measures allowing for evaluating overall model adequacy and model selection, we observe that they tend to select model 3, except P_{rand}, which identifies the population model for the majority of the 200 replications.

In summary, we have seen that (a) the considered measures of overall model adequacy, except λ_y, R^2_α, λ_α, R^2_β and λ_β and those based on the null model 2, have similar sensitivity to the modifications of the model parameters. Moreover, they behave as expected as their values are bigger when the variability in the simulated data is higher; (b) the measures of (Gelman and Pardoe, 2006) are highly sensitive to the modifications of the parameters at their corresponding "level." They thus allow for understanding the relative importance of predictors and error at each "level", as stated by the authors. However, the variability for R^2_α, λ_α, R^2_β and λ_β is very large (cf. supplementary Figure 6). Thereby, it prevents the definition of some guidelines for their use for model selection; (c) the measures based on the null model 2 are, as expected, not sensitive to the modifications of the values of σ_α. And they behave similarly to their counterparts based on the null model 0, but with smaller effects when we manipulate ρ and σ_β. However, the manipulation of γ_1 and δ_1 imply the reverse behavior of their counterparts based on the null model 0. Indeed, bigger values of these parameters give rise to smaller values of the measures. Measures based on the null model 0 thus seem to be preferable to their counterparts based on the null model 2.

Given the results summarized in points (a), (b) and (c), we advise, for the evaluation of overall model adequacy, the use of one of the considered measures except those of (Gelman and Pardoe, 2006) and those based on the null model 2. To further use these measures for model selection, we have seen that (d) P_{rand} was the only measure able to identify the population model. Thus, we advise to use it when the interest of the researcher is both on selecting the model that best fits the data and on testing overall adequacy of the retained model.

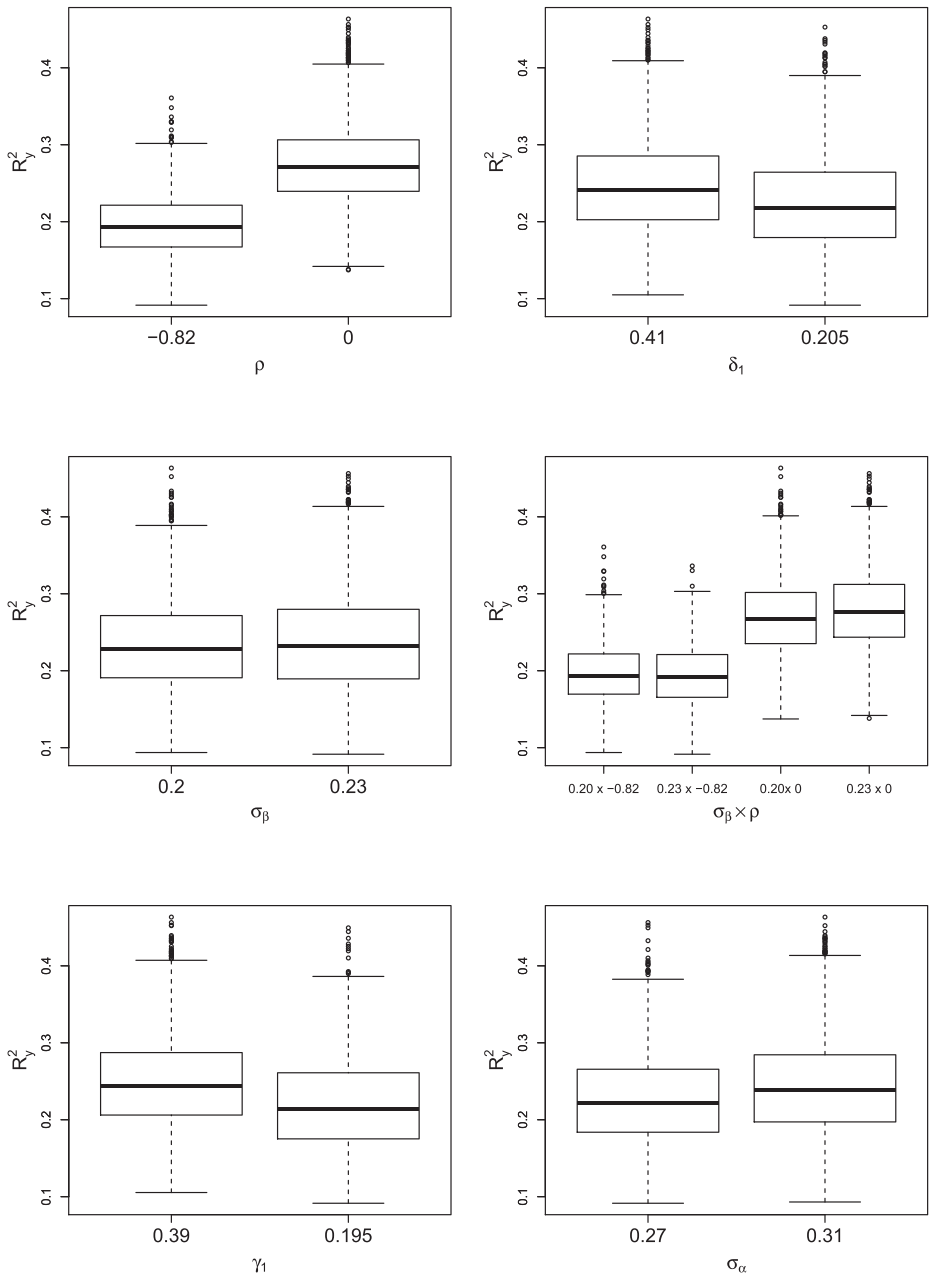

Fig. 1. Boxplots of R_y^2 for the alternative values of parameters considered in the simulation (cf. Table 4) and the subsequent ANOVA analysis.

5.2.3. Comparison of measures of model adequacy due to fixed effects

As for the measures of overall model adequacy, we first discuss the ANOVAs results presented in Table 8. These results are similar for all the measures allowing for fixed effects selection and for evaluating model adequacy due to fixed effects (category F&S). As these measures do not use the estimated random effects in their computation, they are insensitive to the modifications of σ_α and σ_β. Even if σ_α, and sometimes σ_β, are significant in the ANOVAs, the associated partial η^2 values are tiny and we observe superficial differences between boxplots for different values of σ_α and σ_β, respectively (cf. supplementary Figure 7). When the coefficients associated with the predictors γ_1 and δ_1 are bigger, these measures are bigger, as expected. The information between intercept and slope is being pooled when ρ is nonzero, thus the predictors allow for explaining more of the variability in the data and the measures are higher.

We evaluate the ability of the measures in category F&S to select the correct fixed part of the model (cf. Table 9). We expect these measures to select mostly either model 5 or 6 as they are equal in their fixed part. For the majority of the 200 replications, model 5 is selected by all these measures, except the marginal $R^2_{VC,a}$, the marginal c and $R^2_{\mathrm{F},a}$, which select model 4, especially when $\delta_1 = 0.205$.

In order to choose the best set of fixed effects, all the considered measures belonging to category F&S seem to be appropriate, except the marginal $R^2_{VC,a}$, the marginal c and $R^2_{F,a}$. Indeed, they all behave as expected when testing their sensitivity to the modifications of the model parameters and they select the correct fixed part of the model.

5.2.4. Comparison of measures of model selection

First, we computed an ANOVA for each measure as in Sections 5.2.2 and 5.2.3 to evaluate the sensitivity of the measures of model selection (category S). The comments of this paragraph are based on the part of Table 8 not discussed in Sections 5.2.2 and 5.2.3. The sensitivity of -2LL ML, the mAIC and the BIC is identical as they are equal up to an additive constant and -2LL REML provides similar results. For these measures, the parameters of the error structure, ρ, σ_α and σ_β, are significant, but with a tiny partial η^2 (< 0.001) for σ_β. As expected, when ρ is nonzero, the model is more complex and the measures are thus smaller; when σ_α and σ_β are bigger, the measures are also bigger, as the variability in the data is larger (cf. supplementary Figure 8). For the cAIC ML and the cAIC REML, the parameters σ_α and ρ are significant and, as for the mAIC, the cAIC ML and the cAIC REML are smaller when ρ is nonzero and are larger for bigger values of σ_α but to a lesser extent (cf. supplementary Figure 9; similar boxplots are obtained for the cAIC ML). For the DIC, all the main effects are significant except for σ_β. We expect that when the variability in the data increases, the value of the DIC increases as well, which is the case for γ_1 and δ_1. At the opposite, we observe unexpectedly smaller values of the DIC for $\rho = 0$ and bigger σ_α values (cf. supplementary Figure 10). In terms of sensitivity to the modifications of the parameters, considering the conditional likelihood rather than the marginal likelihood is more appropriate for LMMs as the cAIC is less influenced by the modifications of the parameters σ_α and σ_β than the mAIC. Moreover, the cAIC behaves more appropriately than the DIC in this simulation study.

Second, we evaluate the measures in category S in terms of performance in model selection using the systematic rule described in Section 5.2. As the data are simulated from the population model 6, we expect these measures to select mostly this model. For simulation cases 17 to 32, for which the correlation is fixed at zero, either model 5 or 6 are considered as correct. The comments of this paragraph are based on the part of Table 9 not discussed in Sections 5.2.2 and 5.2.3. The information criteria tend to select a model other than the population model, mainly in cases 1 to 16, in which $\rho = -0.820$, except the cAIC REML. The cAIC REML identifies the population model in cases 1 to 16 when the coefficient associated to the cross-level interaction between uranium and basement is higher ($\delta_1 = 0.410$), and in cases 17 to 32 (cf. supplementary Table 1). The smallest values of -2LL ML are always obtained for model 6 as expected. In the odd simulation cases, when $\delta_1 = 0.410$, the smallest values of -2LL REML are obtained for model 6, but for the even cases, when $\delta_1 = 0.205$, they are obtained for model 4. This observation is not problematic because -2LL REML is used to compare the random part of a model. And models 4 and 6 having the same random part and differing only in their fixed part, we would not use -2LL REML to compare these two models.

6. Discussion and Conclusion

In this article we discuss measures allowing evaluating model adequacy and selection in the linear mixed-effects model. In particular, we focus on extensions of R^2 and on information criteria. Thus, we deliberately do not consider penalization approaches for variable selection, such as the adaptive least absolute shrinkage and selection operator (ALASSO; Zou, 2006) or the smoothly clipped absolute deviation (SCAD; Fan and Li, 2001) in the frequentist framework, or the different versions of spike and slab priors (see Section 3.2 of Fahrmeir et al., 2009) in the Bayesian framework. The comparison of such methods is left for future research. Note that, because we were initially motivated by the example of (Gelman and Pardoe, 2006), in our simulation study design the population model 6, defined by Equation (4) was the most complex. Hence, we examined how the various measures fare in rejecting a simpler model in light of the correct and more complex model. We did not address the question of how well these measures reject more complex models (i.e., containing unneeded parameters) when compared to the correct and simpler model.

Based on the results of the simulation study in Section 5.2 and the characteristics of the measures partly listed in Table 1, we now discuss some guidelines for what appear to be the most promising measures. All the measures evaluating overall model adequacy (categories A and A&S), except the measures of (Gelman and Pardoe, 2006) and those based on a null model with a fixed intercept and a random intercept, give similar simulation results in terms of sensitivity to the modifications of some parameters. Some of them have the advantage to be further used for model selection (category A&S) as they include some sort of a penalty function (e.g., adjustment for the number of parameters). Among them, P_{rand} was the only measure able to identify in all simulation cases the population model, certainly due to its penalty term for large random effects. According to our results, we can also favor the use of the conditional $r_{c,a}$ in addition to P_{rand}. Indeed, the conditional $r_{c,a}$ does not require the specification of a null model, does neither assume normality nor constant variances for the random effects and the errors, and is defined for nonlinear mixed-effects models. Future research should also consider the need of normality for the random effects for P_{rand}. Also, extensions to nonlinear mixed-effects models (Davidian and Giltinan, 1995; Vonesh and Chinchilli, 1996) ought to be investigated.

Concerning the measures that allow assessing model adequacy due to fixed effects and identifying the best fixed part of the model (category F&S), they all yield similar results in the simulation study, except the marginal $R^2_{VC,a}$, the marginal c and $R^2_{F,a}$ that fail in identifying the best set of fixed effects. Among the remaining measures that correctly identify the best

fixed part of the model, we favor the use of the marginal P_{rand} and of the marginal $r_{c,a}$, as their conditional counterparts (P_{rand} and the conditional $r_{c,a}$) perform well for testing overall model adequacy and selecting model.

In our simulation, all the information criteria but one wrongly selected a model different from the population model. Only the cAIC REML did not fail in this regard, most probably due to its computation based on the conditional likelihood and to the use of a penalty function adapted for LMMs that is different depending on whether the LMM is estimated with ML or REML.

For model specification, researchers should ideally consider all plausible models and apply some of the measures described in this paper. However, this approach can result in a very large number of alternative models and thus becomes unfeasible. Guidelines on how to proceed in practice for model specification can be found in Chapter 6 of (Snijders and Bosker, 1999). Another interesting topic discussed in Chapter 9 of (Snijders and Bosker, 1999) is the possibility to include contextual effects to model a difference between the within- and between-group regression coefficients of a variable. Treating both model specification and contextual effects was however beyond the scopes of this article.

To conclude, for researchers wanting to compare LMMs and to evaluate their adequacy in order to capture as much information as possible in the data, we recommend to consider jointly a measure allowing for (a) model selection (cAIC REML), (b) fixed effects selection (the marginal $r_{c,a}$ and/or the marginal P_{rand}), and (c) testing overall model quality (the conditional $r_{c,a}$ and/or P_{rand}). The choice between (marginal) P_{rand} or the (marginal) conditional $r_{c,a}$ comes down to the interpretation. The former can be interpreted similarly to the classical R^2, as it measures the proportional reduction in PQL, whereas the latter is less commonly interpreted as a concordance correlation coefficient between observed and predicted values.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ecosta.2021.05.005.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19 (6), 716–723. doi:10.1109/TAC.1974.1100705.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67 (1), 1–48. doi:10.18637/jss.v067.i01.

Bondell, H.D., Krishna, A., Ghosh, S.K., 2010. Joint variable selection for fixed and random effects in linear mixed-effects models. Biometrics 66 (4), 1069–1077. doi:10.1111/j.1541-0420.2010.01391.x.

Bryk, A.S., Raudenbush, S.W., 1992. Hierarchical linear models, 1st Sage.

Burnham, K.P., Anderson, D.R., 2002. Model selection and multimodel inference: A practical information-theoretic approach. New York: Springer.

Chen, Z., Dunson, D.B., 2003. Random effects selection in linear mixed models. Biometrics 59 (4), 762–769. doi:10.1111/j.0006-341X.2003.00089.x.

Cohen, J., 1988. Statistical power analysis for the behavioral sciences, 2nd. Lawrence Elrbaum Associates, Hillsdale, NJ.

Davidian, M., Giltinan, D.M., 1995. Nonlinear models for repeated measurement data. Chapman & Hall Ltd.

Draper, N.R., Smith, H., 1998. Applied regression analysis, 3rd Wiley.

Edwards, L.J., Muller, K.E., Wolfinger, R.D., Qaqish, B.F., Schabenberger, O., 2008. An R^2 statistic for fixed effects in the linear mixed model. Statistics in Medicine 27 (29), 6137–6157. doi:10.1002/sim.3429.

Efron, B., 2004. The Estimation of Prediction Error. Journal of the American Statistical Association 99 (467), 619–632. doi:10.1198/016214504000000692.

Fahrmeir, L., Kneib, T., Konrath, S., 2009. Bayesian regularisation in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection. Statistics and Computing 20 (2), 203–219. doi:10.1007/s11222-009-9158-3.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96 (456), 1348–1360. doi:10.1198/016214501753382273.

Gelman, A., Pardoe, I., 2006. Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. Technometrics 48 (2), 241–251. doi:10.1198/004017005000000517.

Goldstein, H., 2011. Multilevel statistical models, 4th Wiley.

Greven, S., Kneib, T., 2010. On the behaviour of marginal and conditional AIC in linear mixed models. Biometrika 97 (4), 773–789. doi:10.1093/biomet/asq042.

Gruber, L.F., West, M., 2017. Bayesian online variable selection and scalable multivariate volatility forecasting in simultaneous graphical dynamic linear models. Econometrics and Statistics 3, 3–22.

Hodges, J.S., Sargent, D.J., 2001. Counting degrees of freedom in hierarchical and other richly-parameterised models. Biometrika 88 (2), 367–379. doi:10.1093/biomet/88.2.367.

Jiang, J., Rao, J.S., Gu, Z., Nguyen, T., 2008. Fence methods for mixed model selection. The Annals of Statistics 36 (4), 1669–1692. doi:10.1214/07-AOS517.

Kiviet, J.F., 2020. Microeconometric dynamic panel data methods: Model specification and selection issues. Econometrics and Statistics 13, 16–45.

Ko, V., Hjort, N.L., 2019. Copula information criterion for model selection with two-stage maximum likelihood estimation. Econometrics and Statistics 12, 167–180.

Korn, E.L., Simon, R., 1991. Explained residual variation, explained risk, and goodness of fit. The American Statistician 45 (3), 201–206. doi:10.2307/2684290.

Laird, N.M., Ware, J.H., 1982. Random-effects models for longitudinal data. Biometrics 38 (4), 963–974. doi:10.2307/2529876.

Leschinski, C., Sibbertsen, P., 2019. Model order selection in periodic long memory models. Econometrics and Statistics 9, 78–94.

Liang, H., Wu, H., Zou, G., 2008. A note on conditional AIC for linear mixed-effects models. Biometrika 95 (3), 773–778. doi:10.1093/biomet/asn023.

Lin, L.I., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45 (1), 255–268. doi:10.2307/2532051.

Liu, H., Zheng, Y., Shen, J., 2008. Goodness-of-fit measures of R^2 for repeated measures mixed effect models. Journal of Applied Statistics 35 (10), 1081–1092. doi:10.1080/02664760802124422.

Müller, S., Scealy, J., Welsh, A., 2013. Model Selection in Linear Mixed Models. Statistical Science 28, 135–167. doi:10.1214/12-STS410.

Nakagawa, S., Johnson, P.C.D., Schielzeth, H., 2017. The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. Journal of The Royal Society Interface 14 (134), 20170213. doi:10.1098/rsif.2017.0213.

Nakagawa, S., Schielzeth, H., 2013. A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods in Ecology and Evolution 4 (2), 133–142. doi:10.1111/j.2041-210x.2012.00261.x.

Orelien, J.G., Edwards, L.J., 2008. Fixed-effect variable selection in linear mixed models using R^2 statistics. Computational Statistics & Data Analysis 52 (4), 1896–1907. doi:10.1016/j.csda.2007.06.006.

Pan, Z., Lin, D.Y., 2005. Goodness-of-fit methods for generalized linear mixed models. Biometrics 61 (4), 1000–1009. doi:10.1111/j.1541-0420.2005.00365.x.
Pu, W., Niu, X.-F., 2006. Selecting mixed-effects models based on a generalized information criterion. Journal of Multivariate Analysis 97 (3), 733–758. doi:10.1016/j.jmva.2005.05.009.
R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Saefken, B., Ruegamer, D., Greven, S., Kneib, T., 2014. cAIC4: Conditional Akaike information criterion for lme4. R package version 0.2
Sakamoto, W., 2019. Bias-reduced marginal Akaike information criteria based on a Monte Carlo method for linear mixed-effects models. Scandinavian Journal of Statistics 46 (1), 87–115. doi:10.1111/sjos.12339.
Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6 (2), 461–464. doi:10.2307/2958889.
Skrondal, A., Rabe-Hesketh, S., 2004. Generalized latent variable modelling: Multilevel, longitudinal, and structural equation models. Chapman & Hall/CRC.
Snijders, T.A.B., Bosker, R.J., 1994. Modeled variance in two-level models. Sociological Methods & Research 22 (3), 342–363. doi:10.1177/0049124194022003004.
Snijders, T.A.B., Bosker, R.J., 1999. Multilevel analysis: an introduction to basic and advanced multilevel modeling. SAGE Publications, London.
Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 64 (4), 583–639. doi:10.1111/1467-9868.00353.
Stan Development Team, 2018. RStan: the R interface to Stan.
Säfken, B., Kneib, T., 2020. Conditional covariance penalties for mixed models. Scandinavian Journal of Statistics 47 (3), 990–1010. doi:10.1111/sjos.12437.
Vaida, F., Blanchard, S., 2005. Conditional Akaike information for mixed-effects models. Biometrika 92 (2), 351–370. doi:10.1093/biomet/92.2.351.
Vonesh, E.F., Chinchilli, V.M., 1996. Linear and nonlinear models for the analysis of repeated measurements. CRC Press.
Vonesh, E.F., Chinchilli, V.M., Pu, K., 1996. Goodness-of-fit in generalized nonlinear mixed-effects models. Biometrics 52 (2), 572–587. doi:10.2307/2532896.
Wheeler, D., Hickson, D., Waller, L., 2010. Assessing local model adequacy in Bayesian hierarchical models using the partitioned deviance information criterion. Computational statistics & data analysis 54 (6), 1657–1671. doi:10.1016/j.csda.2010.01.025.
Wilks, S.S., 1938. The large-sample distribution of the likelihood ratio for testing composite hypotheses. The Annals of Mathematical Statistics 9 (1), 60–62. doi:10.1214/aoms/1177732360.
Xu, R., 2003. Measuring explained variation in linear mixed effects models. Statistics in Medicine 22 (22), 3527–3541. doi:10.1002/sim.1572.
Zheng, B., 2000. Summarizing the goodness of fit of generalized linear models for longitudinal data. Statistics in Medicine 19 (10), 1265–1275. doi:10.1002/(SICI)1097-0258(20000530)19:10<1265::AID-SIM486)3.0.CO;2-U.
Zou, H., 2006. The adaptive Lasso and its oracle properties. Journal of the American Statistical Association 101 (476), 1418–1429. doi:10.1198/016214506000000735.