# Variable selection in mixed models: a graphical approach

## *Selezione di variabili in modelli misti: un approccio grafico*

Mariangela Sciandra

**Abstract** Model selection can be defined as the task of estimating the performance of different models in order to choose the (approximate) best one. The purpose of this article is to introduce an extension of the graphical representation of deviance proposed in the framework of classical and generalized linear models to the wider class of mixed models. The proposed plot is useful in determining which are the important explanatory variables conditioning on the random effects part. The applicability and the easy interpretation of the graph are illustrated with a real data examples.
**Abstract** *Scopo di questo articolo e' quello di introdurre un'estensione alla rappresentazione grafica della devianza proposta nell'ambito dei modelli lineari alla classe dei modelli lineari misti. La rappresentazione grafica proposta e' particolarmente utile per la determinazione delle variabili esplicative da inserire nel predittore lineare della componente fissa di un modello misto condizionatamente alla parte casuale. Il semplice impiego e la facile interpretazione del grafico sono illustrati con uno esempio su dati reali.*

**Key words:** Graphical representation, Mixed models, Model selection, Penalized Weighted Residual Sum of Squares.

## 1 Introduction

Mixed models (linear and generalized linear) have rapidly become a widely used tool for modelling clustered and longitudinal data. The application of mixed models is vast and expanding so fast as to preclude any attempt at exhaustive coverage. The areas of application include biological and medical research, animal and human genetics, and small area estimation. Although a large amount of work has been done in

Dipartimento di Scienze Economiche Aziendali e Statistiche, e-mail: mariangela.sciandra@unipa.it

the literature on likelihood-based inference for mixed models, model selection for mixed models continues to be a complex and interesting area of research (*Edwards et al., 2008*). For the most part, it is assumed that the basic assumptions about the model, for example, those about the presence of the random effects and their distributions, are correct. However, sometimes it is not clear which is the best model to use when there are a number of potential, or candidate, models. Here being best is in the sense that the model is not only correct but also most economical, meaning that it is simplest among all correct models. In the following the problem of mixed model selection will be considered by a graphical point of view assuming that random effect factors are not subject to selection. In particular, the graph proposed by Brown (1992) in the context of classical linear modelling is generalized to its application for mixed models. The article is organized as follows. The model selection problem in mixed models will be discuss in section 2. The proposed graphical generalization and the real data examples will be illustrated in Section 3.

## 2 Mixed models selection

The notion of decrease in residual variation due to the use of a regression model actually incorporates two concepts: (1) the global capability of the model to reproduce the data generating process and (2) the specific contribution of the covariates to the explanation of the phenomenon. The first concept is generally referred as "goodness of fit" and the second concept as "explained variation". For the linear regression model the usual coefficient of determination $R^2$ incorporates both concepts and represents the widest used measure of model selection. In the linear model $R^2$ corresponds to compare two models: a "full" model that consists of $p-1$ independent predictors in addition to an intercept and a "null" model containing only the intercept. This is no longer true in mixed models where we could be interested in $R^2$ as a general measure of goodness-of-fit of the model to the data rather than a measure of explained variation. This difference from the linear regression case is made worse if we consider that while in fixed-effects models the concept of explained variation refers necessarily to the reduction in variation due to the fixed covariates, in mixed models we have to distinguish between variation explained by the fixed part, variation explained by the random effects and variation explained through a properly assumed correlation structure. So, in order to evaluate the proportion of explained variation a primary step has to define which model has to be meant as "null" model. According to the different situations we can consider as "null" model the model with both fixed and random intercepts or, in cases in which the clustering variable is itself of interest, the "null" model could no longer include the random intercept. Aim of this paper is to focus on a graphical comparison of several mixed models that only differ for the fixed part. So in building the plot it is necessary to use a measure that allows to express quantitatively the gain in the proportion of residual variability explained by the inclusion of an extra fixed component into the fixed part of the model conditioning to the same random part and covariance structure. In the

following subsection a brief recall to Linear Mixed Model (LMM) and Generalized Linear Mixed Models (GLMM) definitions will be given and the model selection measures properly derived.

## 2.1 Measures of model selection in LMM and GLMM

According to Laird and Ware (1982), the general Linear Mixed Model can be summarized as follows:

$$\begin{cases} \boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i \\ \boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}), \\ \boldsymbol{\varepsilon}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_i), \\ \boldsymbol{b}_1, \ldots, \boldsymbol{b}_m, \boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_m \quad \text{independent,} \end{cases} \tag{1}$$

where $\boldsymbol{Y}_i$ is the $n_i$-dimensional response vector for subject $i$, $1 \leq i \leq m$, $m$ is the number of subjects, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are $(n_i \times p)$ and $(n_i \times q)$ dimensional matrices of known covariates, $\boldsymbol{\beta}$ is a $p$-dimensional vector containing the fixed effects, $\boldsymbol{b}_i$ is the $q$-dimensional vector containing the random effects and $\boldsymbol{\varepsilon}_i$ is a $n_i$-dimensional vector of residual components. Finally, $\boldsymbol{D}$ is a general $(q \times q)$ covariance matrix and $\boldsymbol{\Sigma}_i$ is a $(n_i \times n_i)$ covariance matrix which depends on $i$ only through its dimension $n_i$, i.e. the set of unknown parameters in $\boldsymbol{\Sigma}_i$ does not depend upon $i$. The marginal density function of $\boldsymbol{Y}_i$ is then an $n_i$ dimensional Normal distribution with mean vector $\boldsymbol{X}_i\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \boldsymbol{\Sigma}_i$. The $\boldsymbol{V}_i$ structure is assumed to be the same for all subjects, but we continue to use the subscript $i$ to emphasize that we are referring to the covariance matrix for a single subject. Observations on different subjects are assumed to be independent. Several measures of explained variation for the gaussian linear mixed model case have been proposed in literature (*Xu,2003*, *Edwards* et al.,*2008*, *Hossjer, 2008*). In this work the model selection problem will consist into selecting the fixed covariates from a set of candidate predictors when the random effects part of the model (i.e. $\boldsymbol{Z}_i\boldsymbol{b}_i$) is not subject to selection. According to Fitzmaurice, Laird, and Ware (2007), twice the difference between the maximized log-likelihoods of two nested models, $G^2 = 2(\hat{l}_{\texttt{full}} - \hat{l}_{\texttt{reduced}})$, represents the degree to which the reduced model is inadequate. As the aim of this work is to compare mean models with the same covariance structure it is possible to define:

1. Model of interest ("*full*"): $\boldsymbol{Y}_i = \boldsymbol{F}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i$
2. Reduced Model ("*null*"): $\boldsymbol{Y}_i = \boldsymbol{R}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i$

where $\boldsymbol{F}_i$ is a $n_i \times p$ matrix of known fixed covariates and $\boldsymbol{R}_i$ is an its submatrix with the same number of rows but with a reduced number of columns. When the sample size is large, $G^2 \sim \chi^2$ with degrees of freedom equal to the difference in parameters between the "full" and reduced models. The use of such a test is a quite general principle for statistical testing. The deviance can be regarded as a measure of lack of fit between model and data. In general, the larger is the deviance, the poorer is the fit to the data. In a GLMM context it is assumed that, conditionally on the $q$-

dimensional random effects $\mathbf{b}_i$, the elements $Y_{ij}$ of $\mathbf{Y}_i$ are independent and follow a distribution in the exponential family:

$$f\{Y_{ij}|\boldsymbol{b}_i,\beta,\phi\} = exp\{\phi^{-1}[y_{ij}\vartheta_{ij} - \psi(\vartheta_{ij})] + c(y_{ij},\phi)\} \quad j = 1,\ldots,n_i \tag{2}$$

with $\mu_{ij} = E(Y_{ij}|\boldsymbol{b}_i) = \psi'(\vartheta_{ij})$ and $Var(Y_{ij}|\boldsymbol{b}_i) = \psi''(\vartheta_{ij})\phi$, $Var(Y_{ij}|\boldsymbol{b}_i) = \psi''(\vartheta_{ij})\phi$, where $\vartheta_{ij}$ is the canonical parameter; finally, a known link function $g(\cdot)$ relates the linear predictor to the transformed mean response as

$$g(\mu_{ij}) = g[E(Y_{ij}|\boldsymbol{b}_i)] = \boldsymbol{x}_{ij}^T\beta + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i \tag{3}$$

with $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ the corresponding $(p \times 1)$ and $(q \times 1)$ vectors of covariates associated with the fixed effects and random effects, respectively, and $\phi$ a dispersion parameter. Finally, in addition to the distributional assumption (2), the random effects are assumed to be drawn independently from $\mathcal{N}(\mathbf{0},\mathbf{D}(\tau))$. The presence of random parameters prevents the use of a standard likelihood function and hence a remarkable amount of work in the literature on likelihood-based inference for GLMMs has been devoted to the search of extended versions of the likelihood function and of the deviance measures of discrepancy commonly used in standard GLMs (for a review, see *Molenberghs and Verbeke, 2005*). In order to generalize the deviance plot to the GLMM cases, a measure of discrepancy between models that depends on both the fixed parameters $\beta,\tau$ and on the random parameters $\mathbf{b}_i, i = 1,\ldots,m$ and having the property of being monotonic over the lattice indexing all possible models has to be defined. In this work the **Penalized Weighted Residual Sum of Squares (PWRSS)** introduced by Bates(2013) will be used. It is defined as:

$$PWRSS = [\mathbf{y} - \mathbf{g}^1(\mathbf{Z}\Lambda(\tau)\mathbf{u} + \mathbf{X}\beta)]^T\mathbf{W}[\mathbf{y} - \mathbf{g}^1(\mathbf{Z}\Lambda(\tau)\mathbf{u} + \mathbf{X}\beta)] + \parallel \mathbf{u} \parallel \tag{4}$$

where: $\mathbf{y} = [\mathbf{y}_1^T,\ldots,\mathbf{y}_m^T]^T$, $\mathbf{X} = [\mathbf{X}_1^T,\ldots,\mathbf{X}_m^T]^T$, $\mathbf{Z} = \bigoplus_{i=1}^m \mathbf{Z}_i$, $\mathbf{W} = \bigoplus_{i=1}^m \mathbf{W}_i$, with $\mathbf{W}_i$ the usual weight matrix of the conditional GLM in the $i$-th cluster, $\mathbf{b}_i = \Lambda(\tau)\mathbf{u}_i, i = 1,\ldots,m$, and $\mathbf{u} = [\mathbf{u}_1^T,\ldots,\mathbf{u}_m^T]^T$, with $\Lambda(\tau)$ a transformation matrix such that $\mathbf{u} \sim \mathcal{N}(\mathbf{0},\mathbf{I}_n)$.

## 3 Graphical selection tool

In standard linear regression it is a common practice to select a subset of variables from a complex model by using a stepwise regression procedure in which the impact of each variable to the model fitting is evaluated step by step. There are several versions of stepwise regression such as forward selection, backward elimination, and stepwise. Many researchers employed these techniques to determine the order of predictors by its magnitude of influence on the outcome variable. However, the above interpretation is valid if and only if all predictors are independent. When regressors will show some degree of correlation such procedures would return inaccurate results. In such situations, between several non-orthogonal regressors, the

one entering the model first will seem to contribute the largest amount of explained variance. The remaining predictors will seem to be less influential because their contribution to the explained variance will be overlapped by the variable entering the model first. Indeed, the more correlated the regressors are, the more their ranked "importance" will depend on the selection order. A valid alternative consists in evaluating all the possible $2^p$ submodels and summarizing in a table the changes observed in some measures of goodness of fit occurring when terms are added or dropped out of the model. Brown (1992) in his paper suggested a graphical representation useful for the assessment of results and the comparison of all possible linear and generalized linear models that is possible to fit with $p$ regressors. In this paper such graphical representation is generalized to the class of mixed models in order to select fixed regressors conditioning on the same random part and covariance structure. In order to present the method we will make use of a simple example. Let us consider toy data where the outcome variable *Extraversion* (Extra) is predicted by fixed effects for the interval scaled predictor *Openness to new experiences* (A), the interval scaled predictor *Agreeableness* (B), the interval scaled predictor *Social engagement* (C); as well as the random (nested) effect of the nominal scaled predictor *Class* within the nominal scaled predictor *School*. The data contains 1200 cases evenly distributed among 24 nested groups (4 classes within 6 schools). The selection of the variable entering the fixed linear predictor is carried out graphically using the plot in Fig 1.

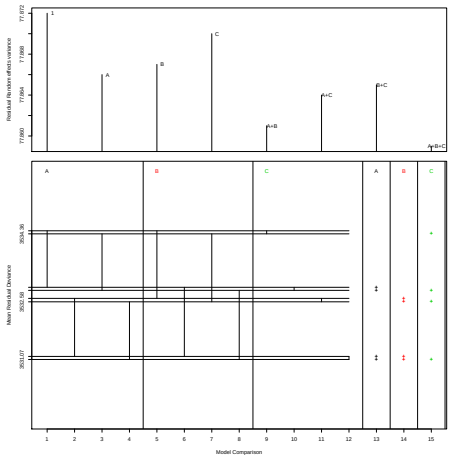


**Fig. 1** *Graphical analysis of PWRR for the* Extraversion *data.*

The graph is constituted by two frames. The lowest frame on the x-axis reports the number of model comparisons ($2^p$) that can be done using $p$ regressors; on the vertical axis the marginal deviances associated to each models are reported where each model is represented by a horizontal line in the left-hand side of the graph. The lower right-hand side of the graph contains a numbers of columns equal to

the number of possible predictors. The presence of a specific regressor in the linear predictor of the fixed par of the fixed model is indicated by a plus sign in the appropriate column. When the column is empty this means that the term is not included in that specific model. So, the "null" model, the one including only the fixed intercept will have all blank columns and will be the one with the highest residual deviance associated. In the example of *Extraversion* we fitted several linear mixed effect models with several fixed component specification, and a random nested effect of class within school. Looking each box separately it is possible to derive the reduction in PWRSS observed when the variable indexing the box is added. So, in the first box for example, the vertical lines give the reduction in PWRSS observed when the predictor A is added to different models; the second box includes comparison between models in which the predictor B is added and so on. We can plot a number of boxes equal to the number of fixed predictors ($p$). In the example above, then, the variable C is the less important because to its introduction are associated the smallest reductions in PWRSS. The greatest reduction is the one occurred when the predictor B is added to the model containing A. So, there are in all $2^p$ horizontal lines representing the fitted models and each of them is connected with all possible more parsimonious nested models by vertical lines. The lengths of this vertical lines represent the reduction in PWRSS associated to the addition of a term in the fixed linear predictor. A critical analysis of the plot allows to select the best fixed effect part specification keeping in mind that all the compared models are equal defined in their random effects parts. In order to have a complete overview of the models that we are comparing, in the highest frame of the plot the values of the residual variances of the random effect part associated with the different fixed part specified models are reported. So, for example, in the example about Extraversion we can conclude that, when the fixed component includes additively both regressors A and B, the residual variance is subject to an important reduction if compared with all the models containing a single fixed component.

## References


1. Bates, D.: Computational methods for mixed models. *http://cran.r-project.org/web/packages/lme4/vignettes/Theory.pdf*.
2. Brown, D.: A Graphical Analysis of Deviance. Journal of the Royal Statistical Society. Series C (Applied Statistics), **41**, 55–62 (1992)
3. Fitzmaurice, G.M., Laird, N.M., Ware, J.H.: Applied Longitudinal Analysis. Wiley, Wiley Series in Probability and Statistics (2007)
4. Hössjer, O.: On the coefficient of determination for mixed regression models. Journal of Statistical Planning and Inference, **38**, 3022–3038 (2008)
5. Laird, N.M., Ware, J.H.: Random-Effects Models for Longitudinal Data. Biometrics, **38**, 963–974 (1982)
6. Molenberghs,G., Verbeke,G.: Models for Discrete Longitudinal Data. Springer, (2005)
7. Orelien, J.G, Edwards, L.J.: Fixed-effect variable selection in linear mixed models using $R^2$ statistics. Computational Statistics & Data Analysis, **52**, 1896–1907 (2008)
8. Xu, R.: Measuring explained variation in linear mixed effects models. Statistics in Medicine, **22**, 3527–3541 (2003)