On Generalized Degrees of Freedom and their Application in Linear Mixed Model Selection

Chong You[1,2], Samuel Müller[1], John T. Ormerod[1]

[1]School of Mathematics and Statistics, University of Sydney, NSW 2006 Australia

[2]Corresponding author: Chong You, e-mail: C.you@maths.usyd.edu.au

Abstract

The concept of degrees of freedom plays an important role in statistical modeling and is commonly used for measuring model complexity. The nominal degrees of freedom, i.e., the number of unknown parameters, may fail to work in some modeling procedures, in particular linear mixed effects model situations. In this article, we proposed a new definition for generalized degrees of freedom in the context of linear mixed effects models. It is derived based on the sum of the sensitivity of the expected fitted values with respect to their underlying true means and can be simplified to a sum of covariance terms. Furthermore, we explore and compare two estimation methods, data perturbation and residual bootstrap to approximate the proposed generalized degrees of freedom. We also show that this generalized degrees of freedom satisfies some desirable properties and can be used for linear mixed effects model selection.

Keywords: Deviance, Information Criterion, Resampling, Bootstrap

1. Introduction

The concept of degrees of freedom (df) plays an important role in statistical modeling. The nominal degrees of freedom, i.e., the number of unknown parameters, is commonly used in linear regression model for measuring model complexity, particulary when using maximum likelihood or least square estimators for the regression parameters. However, many linear model estimates are not based on these methods. As a consequence the nominal degree of freedom may fail to work. In pioneering work, Ye (1998) proposed a new definition to measure the complexity of a linear model, the generalized degrees of freedom, which is based on the sum of the sensitivity of the expected fitted values to the corresponding expected values of the observations. Ye's generalized df has been widely accepted in linear models. Examples include Lian (2003), Zou et al. (2007) and Vaiter et al. (2012). The linear mixed effects model is an increasing commonly used statistical model and is much more complicated than the simple linear regression model. To our knowledge, very little work on the df in linear mixed effects models has been done. A notable exception is Zhang et al. (2012), who derived a generalized df based on the expected Kullback-Leibler loss and who additionally proposed an adaptive model selection procedure for linear mixed effects models. The generalized df in Zhang et al. (2012) is defined as,

$$gdf_{\text{zhang}} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i} \text{Cov}\left(\hat{\mathbf{V}}_{ijk}(\mathbf{Y})\hat{g}_{ij}(\mathbf{Y}), Y_{ik}\right) - \frac{1}{2}\text{Cov}\left(\hat{\mathbf{V}}_{ijk}(\mathbf{Y}), Y_{ij}Y_{ik}\right), \tag{1}$$

where $\hat{g}_{ij}(\mathbf{Y})$ is a function of $\mathbf{Y}$ which gives the fitted value of Y_{ij} and $\hat{\mathbf{V}}_{ijk}$ is the (j,k)th entry of the inverse of $\hat{\Sigma}_i$.

In this paper, we extend Ye's idea and propose a new definition of generalized df in the linear mixed effects model context. Furthermore, we approximate our generalized df through a residual bootstrap. We show that our proposed generalized df satisfies some desirable properties and that it can be used for the selection of linear mixed effects models.
In the following we focus on a special case of the linear mixed effects model, the cluster model, which can be expressed as

$$\mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \varepsilon_i, \quad i = 1,2,\ldots,m, \tag{2}$$

where $\mathbf{X}_i \in mat(n_i, p)$ and $\mathbf{Z}_i \in mat(n_i, q)$ are the design matrices for the fixed part and the random part respectively, m is the number of clusters, β is a vector of fixed effects, $\mathbf{b}_i$ is a vector of random effects such that $\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$ for all i. The within-group error terms ε_i is independent to the random effects $\mathbf{b}_i$. If not otherwise indicated, we assume that the components of ε_i are independent and $\varepsilon_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_{n_i})$, where n_i is the size of the ith cluster. Thus we consider the independent cluster model (Müller et al., 2013). Hence, model (2) simplifies to $\mathbf{Y}_i \sim \mathcal{N}\left(\mu_i, \Sigma_i\right)$, where $\mu_i = \mathbf{X}_i\beta$ and $\Sigma_i = \sigma^2\mathbf{I}_{n_i} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i'$.
The rest of this paper is organized as follows. Section 2 presents a new definition of generalized df and two ways to estimate it. In Section 3, we describe how the proposed generalized df can be used for linear mixed effect model selection. A simulation study of linear mixed effects model selection is conducted in Section 4. We conclude in Section 5.

2. Generalized Degree of Freedom and its Estimating Methods

2.1. Definition

Motivated from Ye (1998) and Zhang et al. (2012) we propose a new definition of generalized df for the linear mixed effects model in Equation (2). It measures the model complexity according to the sensitivity of the estimated value of Y_{ij} with respect to the corresponding underlying true means. That is,

$$gdf_s = \sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{\partial \mathbb{E}[\hat{g}_{ij}(\mathbf{Y})]}{\partial \mu_{ij}} \tag{3}$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n_i} \int \hat{g}_{ij}(\mathbf{y}) f(\mathbf{y}) \sum_{k=1}^{n_i} \mathbf{V}_{ijk}(y_{ik} - \mu_{ik}) \mathrm{d}\mathbf{y} = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i} \mathbf{V}_{ijk}\mathrm{Cov}(\hat{g}_{ij}(\mathbf{Y}), Y_{ik}), \tag{4}$$

where $\mathbf{V}_{ijk}$ is the (j,k)th entry of the inverse of Σ_i, hence $\mathbf{V}_i$ is symmetric, $\mathbf{V}_{ijk} = \mathbf{V}_{ikj}$.

2.2. Properties of gdf_s

To be a reasonable and consistent measure of model complexity, we suggest that any definition of df should satisfy the following properties: (1) df is non-negative, (2) if model $\mathcal{M}_1$ is nested in model $\mathcal{M}_2$, then df of $\mathcal{M}_2$ should be larger than df of $\mathcal{M}_1$, (3) df equals the nominal df namely the number of unknown parameters if least square estimates are used in linear regression. In You et al. (2013) we provide empirical evidence that our proposed generalized df is non-negative. The rationale is that with the increasing underlying mean of Y_{ij}, reasonable estimates $\hat{g}_{ij}$ should increase as well, and so the derivative in Equation

(3) should be positive. In You et al. (2013) we prove the following lemma that shows that gdf_s coincides with the nominal df if the least square estimate is used in a linear regression model.

Lemma 1: *If the fitted values of the response is* $\hat{\mathbf{g}} = \mathbf{HY}$, *where* $\mathbf{H}$ *is the hat matrix, then* $gdf_s = \mathrm{tr}(\mathbf{H})$. *Especially, when the least square estimates are used,* $gdf_s = \mathrm{tr}(\mathbf{H}) = \mathrm{tr}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) = p$, *the number of unknown parameters.*

2.3. Estimation

From Equation (4) we can see that both $\mathbf{V}_{ijk}$ and the covariance term in gdf_s are unknown. According to the consistency property of the REML estimates Jiang (1996, 1998), we suggest to use the REML estimate of the covariance matrix from the full model to replace $\mathbf{V}_{ijk}$ in Equation 4. The term $\mathrm{Cov}(\hat{g}_{ij}(\mathbf{Y}), Y_{ik})$ can be approximated from the sample covariance which requests to resample $\mathbf{Y}$. The residual bootstrap is a straightforward method to resample $\mathbf{Y}$ and there are multiple ways to implement it. We will use the full model REML estimates to resample $\mathbf{Y}$. Hence the pseudodata are generated through $\mathbf{Y}^* = \hat{\mu}_{\text{BLUE,full}} + \mathbf{r}, \mathbf{r} \sim \mathcal{N}(\mathbf{0}, \hat{\Sigma}_{\text{REML,full}})$. You et al. (2013) show that this residual bootstrap method works much better than using the data perturbation approach suggested in Zhang et al. (2012).

3. Model Selection

In practice, model selection is a key aspect in statistical analysis. The literature on selection of linear mixed effects models has grown rapidly in the last decade and we refer to Müller et al. (2013) for a review. However there is not yet consensus in the statistical community on what model selection method to use. In this section, we introduce a new model selection approach using our proposed generalized df for linear mixed effects models.
Consider a model selection criteria

$$-\log p(\mathbf{y}|\hat{\theta}_{\mathscr{M}}) + \kappa\Delta(\mathscr{M}), \quad \kappa \in (0,\infty),$$

where κ is a penalty multiplier and $\Delta(\mathscr{M})$ is a model complexity term. We suggest to use gdf_s to measure the model complexity. To pick the best model for a fixed κ, the criteria of all the potential models should be calculated and the one with smallest value is considered as the best model. However, in practice it is impossible to go through all models, especially when the model size is large as the estimation of the proposed generalized df is quite time consuming. Motivated by Zhang et al. (2012) we propose a model selection procedure as follows: First, select the best model $\hat{\mathscr{M}}_\lambda$ from all candidate models to minimize the following criteria for each possible λ,

$$-\log p(\mathbf{y}|\hat{\theta}_{\mathscr{M}}) + \lambda|\mathscr{M}|, \lambda \in (0,\infty),$$

where $|\mathscr{M}|$ is the number of independent parameters in model $\mathscr{M}$. Next, the optimal λ is selected by

$$\hat{\lambda} = \underset{\lambda}{\mathrm{argmin}} -\log p(\mathbf{y}|\hat{\theta}_{\hat{\mathscr{M}}_\lambda}) + \kappa \times \widehat{gdf}(\hat{\mathscr{M}}_\lambda), \tag{5}$$

and hence the optimal model is $\hat{\mathscr{M}}_{\hat{\lambda}}$. Although the ideas behind are different, the proposed selection procedure for $\kappa = 1$ fixed is very similar to the one in Zhang et al. (2012).
In Section 4, we compare our method to the adaptive model selection approach in Zhang et al. (2012) when $\kappa = 1$ in both methods.

4. Simulation

In this section we provide numerical illustrations on the performance of the proposed method to select models and compare it to both Zhang et al. (2012)'s adaptive model selection approach and the AIC with nominal df. Consider the following linear mixed effects model

$$Y_{ij} = \alpha + \sum_{k=1}^{4} x_{ijk}\beta_k + \sum_{l=1}^{4} z_{ijl}b_{il} + \varepsilon_{ij}, i = 1,\ldots,n_i,\ j = 1,\ldots,m, \tag{6}$$

where $n_i = 10$, $m = 50$, $\varepsilon_{ij} \sim \mathscr{N}(0,\sigma^2)$ is the error term, $\alpha = 1$ is the intercept, $\mathbf{b}_i = (b_{i1},...,b_{i4}) \sim \mathscr{N}(\mathbf{0},\mathbf{D})$ is the random effect, $\mathbf{D}$ is the covariance matrix with the (j,k)th element equal to $\rho^{|j-k|}$. The elements of the design matrices $\mathbf{X}$ and $\mathbf{Z}$ are generated from the standard normal distribution, the response variable $\mathbf{Y}$ is generated via Equation (6) with $\sigma^2 = 1$, $\rho = 0.5$, $\beta_k = 0,1$ and b_{il} included or excluded from the model. Five different settings are considered for both β and $\mathbf{b}$. Let $c = 0,1,2,3,4$ and $d = 0,1,2$ denote the number of fixed and random effects in the data generating model respectively, i.e., $c = 2$ means the first 2 values of $(\beta_1,...,\beta_4)$ are selected to be 1 and 0 otherwise, $d = 2$ means only b_1 and b_1 are in the model. Note, we use the *lmer* function in **R** to calculate the fitted value $\hat{\mathbf{g}}$ in this simulation.
There are 15 data generating models with all the combination of c and d. Simulations are repeated 100 times for each setting. The average sums of the false positives and false negatives of the above selection procedures are shown in Figure 1. The proposed procedure using gdf_s performs well in general. Zhang et al. (2012)'s approach fails to work when the generating model are linear models but has a very similar performance to AIC when the random effects are involved in the generating model. All three methods work better when the model is larger. The generating models here are fairly simple, results for more complex models are shown in You et al. (2013).

5. Conclusion

This article extends Ye (1998)'s idea and develops a new concept of generalized df in linear mixed effects model. We have numerically shown that the proposed gdf_s satisfies desirable properties, for example, always positive and monotonic increasing in the nested models. Furthermore, the generalized df and the estimation method in Zhang et al. (2012) are compared to what we proposed. As seen from the simulation results, our residual bootstrap method may perform better than the data perturbation method proposed in Zhang et al. (2012) and the ability of selecting models for our generalized df is slightly better than the generalized df in Zhang et al. (2012). Our results also give some motivation for

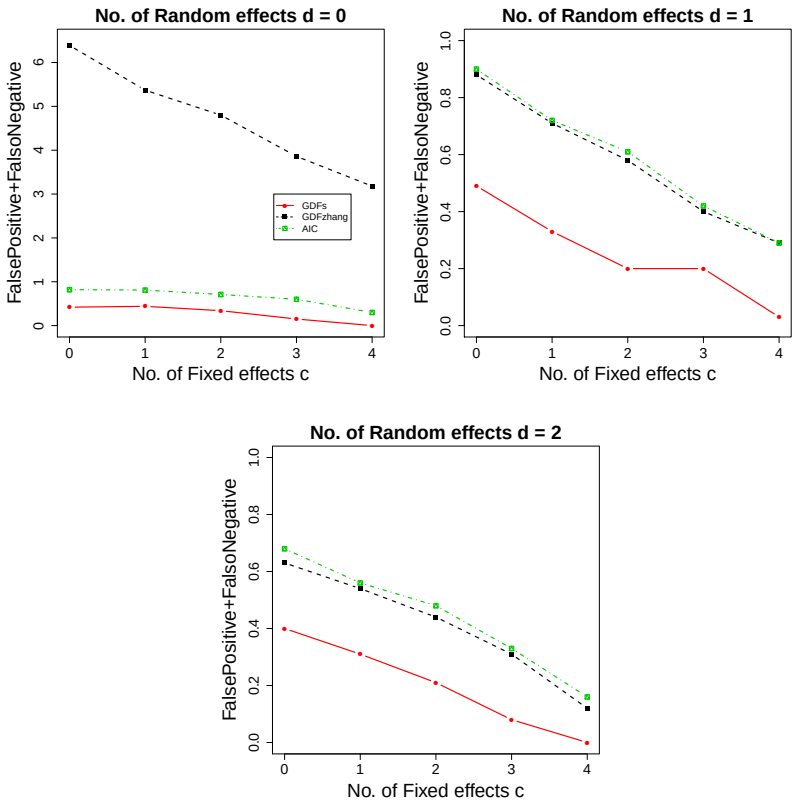

Figure 1: Model selection using gdf_{zhang}, gdf_{s} and AIC

justification of gdf_{s} defined in Equation (3) for other models, such as generalized linear mixed models.

References

Jiang, J., 1996. REML estimation: asymptotic behavior and related topics. Ann. Statist. 24 (1), 255–286.

Jiang, J., 1998. Asymptotic properties of the empirical blup and blue in mixed linear models. Statist. Sinica 8 (3), 861–885.

Lian, I.-B., 2003. Reducing over-dispersion by generalized degree of freedom and propensity score. Comput. Statist. Data Anal. 43 (2), 197–214.

Müller, S., Scealy, J. L., Welsh, A. H., 2013. Model selection in linear mixed models. Statistical Science, 64 pages.

Vaiter, S., Deledalle, C., Peyré, G., Fadili, J., Dossal, C., 2012. The degrees of freedom of the group lasso for a general design. Tech. rep., Preprint Hal-00768896.

Ye, J., 1998. On measuring and correcting the effects of data mining and model selection. J. Amer. Statist. Assoc. 93 (441), 120–131.

You, C., Müller, S., Ormerod, J. T., 2013. On generalized degrees of freedom and their application in linear mixed model selection. Preprint.

Zhang, B., Shen, X., Mumford, S. L., 2012. Generalized degrees of freedom and adaptive model selection in linear mixed-effects models. Comput. Statist. Data Anal. 56 (3), 574–586.

Zou, H., Hastie, T., Tibshirani, R., 2007. On the "degrees of freedom" of the lasso. Ann. Statist. 35 (5), 2173–2192.