Journal of Applied Mathematics & Bioinformatics, vol.2, no.2, 2012, 23-34
ISSN: 1792-6602 (print), 1792-6939 (online)
Scienpress Ltd, 2012

Different Algorithms In Mixed Effect Models Estimation Approach

Manoochehr Babanezhad[1]

Abstract

Standard ordinary linear model cannot handle grouping structure data. This induces a correlation structure through the error term in this model. Therefore mixed effect models often allow the modeling of such structural data setting. This study proposes different estimation algorithms in linear and nonlinear mixed effects models when the grouping structure data are available. I in fact prove a consistency and oracle optimality result in these grounds, and develop algorithms with provable numerical convergence. Further, I demonstrate the performance of the different proposed algorithms on the Landsat ETM+ scene data set.

Mathematics Subject Classification : 62M10
Keywords: Mixed effect, Random effect, Multilevel effect, Tree regression

1 Introduction

Regression analysis is an approach to modeling the relationship between a (scalar or vector) dependent variable and one or more explanatory variables.

[1] Department of Statistics, Faculty of Sciences, Golestan University, Gorgan, Golestan, Iran, e-mail: m.babanezhad@gu.ac.ir.

Article Info: *Received* : June 14, 2012. *Revised* : August 3, 2012
Published online : September 15, 2012

The case of one explanatory variable with linear relationship with the dependent variable is called linear regression model. More than one explanatory variable is called multiple linear regression models. The primary statistical tools are often single-outcome, logistic regression for binary outcomes, and the Cox proportional hazards model for time-to-event outcomes [1, 2]. This group of regression models cannot handle the analysis when one faces on grouping structural data [1, 2, 3]. Instead, mixed effect models have recently come into widespread use. Applying these methods and interpreting the results requires some information [3, 4, 5].

This paper tends to have either very brief coverage or to pay off a single topic theoretically and practically. This study is in fact aimed to show how the combination of common linear and linear mixed effects models can open a huge variety of statistical methods for the analysis of high dimensional data and enrich the inference. The term mixed model refers to the use of both fixed and random effects in the same analysis [5, 6, 7]. Fixed effects have in fact levels that are of primary interest and would be used again if the experiment were repeated [2, 7]. Random effects have levels that are not of primary interest, but rather are thought of as a random selection from a much larger set of levels [7, 8, 9]. For instance in a case study, subject effects are almost always random effects, while treatment levels are almost always fixed effects.

The interest is also in inferences from model-based estimators refer to the distribution implied by the assumed model. That is; model selection and validation play a vital role in this type of estimation [3, 7, 9]. If the assumed models do not provide a good fit to the data, the model-based estimators will be model biased which can lead to (sometimes completely) erroneous inferences [1, 5]. In view of this, this study is summarized in following sections.

In next Section, I explain fixed and random effects. In Section 3 multilevel linear mixed effect model is postulated. The Landsat ETM+ scene data set is analyzed in by taking into account fixed and random effects in Section 4, and conclusion is discussed in Section 5.

2 Fixed and Random Effects

Linear regression model assumes that the relationship between the dependent variable Y and the regressors X is linear. The formula for the linear regression is simple and has interpretable parameters,

$$Y = \beta^{'}\mathbf{X} + \epsilon \tag{1}$$

where β is parameter vector, and $\mathbf{X}$ is matrix of independent variable, and the vector error term ϵ is independently and identically distributed and often has a normal distribution with mean zero and constant variance σ^2 ($\epsilon \sim N(0, \sigma^2 I)$).

The main purpose of this model is to determine how the average value of the continuous outcome Y_i varies with the value of predictor $\mathbf{X}$. The average values of the outcome are assumed to lie on a regression line or line of means. It follows from the model (1) the outcome Y_i has a normal distribution; but often no distributional assumptions are made about the predictor $\mathbf{X}$ in the linear regression model. Although often one does not need to make assumptions about the distribution of the predictor, these models do perform better when it is relatively variable. β gives the slope of the regression line, and is interpretable as the change in average $\mathbf{X}$ for a one unit increase (decrease) in Y.

The regression analysis sometimes designates a type of model that relates a continuous response to both a classification factor and to a continuous regressor. If for instance Y_{ij} is the jth observation in the ith group of data and X_{ij} is the corresponding value of the regressor, in contrast to the model (1), a model with fixed effects β_j for the type factor and random effects b_i for the subject factor can be written [1, 3],

$$Y_{ij} = \beta_j + b_i + \epsilon_{ij} \tag{2}$$

where e.g, $b_i \sim N(0, \sigma_b^2)$ and $\epsilon_{ij} \sim N(0, \sigma^2)$, or, equivalently

$$Y_i = \mathbf{X_i}\beta + Z_i b_i + \epsilon_i \tag{3}$$

where e.g, $b_i \sim N(0, \sigma_b^2)$ and $\epsilon_i \sim N(0, \sigma^2 I)$. The latter model β represents what would be the mean of Y to arise from the jth type of X_{ij} if the whole population were tested. The term analysis of covariance designates a type of model that relates a continuous response to both a classification factor and to a continuous covariate. If y_{ij} is the jth observation in the ith group of data and X_{ij} is the corresponding value of the covariate, an analysis of covariance model with a random effect for the intercept will be model (2). Although many common statistical models can be expressed as linear model. This incorporates with a random effect with individual experimental units drawn at random from a population. Here a model with both fixed effects and random effects is needed. In model (3) , the columns of Z_i are usually a subset of the columns of X_i. Further, the assumption $Var(\epsilon_i) = \sigma^2 I$ can be relaxed in model (3). This can be nonconstant variances or special within-group correlation structures. Furthermore, the random effects bi and the within-group errors ϵ_i, $i = 1, ..., M$ are assumed to be independent for different groups and to be independent of each other for the same group. Because the distribution of the random effects vectors bi is assumed to be normal (or Gaussian) with a mean of 0, it is completely characterized by its variance-covariance matrix which must be symmetric and positive semi-definite; that is, all its Eigen values must be non-negative.

3 Multilevel Linear Mixed Effect Model

Single-level linear mixed effect model (3) can be extended to multiple, nested levels of random effects. For instance, in the case of two nested levels of random effects the response vectors at the innermost level of grouping are written y_{ij}, $i = 1, ..., M$, $j = 1, ..., M_i$ where M is the number of first-level groups and M_i is the number of second-level groups within first-level group i. The fixed-effects model matrices are X_{ij}, $i = 1, ..., M$, $j = 1, ..., M_i$ of size $n_{ij}p$. Using first-level random effects b_i of length q_1 and second-level random effects b_{ij} of length q_2 with corresponding model matrices $Z_{i,j}$ of size $n_i q_1$ and Z_{ij} of size $n_i q_2$, we can postulate the model as follows [1, 6],

$$y_{ij} = \beta X_{ij} + Z_{i,j} b_i + Z_{ij} b_{ij} + \epsilon_{ij} \tag{4}$$

where $b_i \sim N(0, \phi_1)$, $b_{ij} \sim N(0, \phi_2)$, and $\epsilon_{ij} \sim N(0, \sigma^2 I)$. The level-1 random effects b_i are assumed to be independent for different i, the level-2 random effects b_{ij} are assumed to be independent for different i or j and to be independent of the level-1 random effects, and the within group errors ϵ_{ij} are assumed to be independent for different i or j and to be independent of the random effects.

Several methods of parameter estimation have been used for mixed effects models (2), (3), (4) [6, 7]. Although the approaches are different in these three models one often employs on two general methods: maximum likelihood (ML) and restricted maximum likelihood (REML). Consider first the model (2) that has a single level of random effects. The parameters of the model are β, σ^2. We use θ to represent an unconstrained set of parameters. The likelihood function for the model (2) is the probability density for the data given the parameters, but regarded as a function of the parameters with the data fixed, instead of as a function of the data with the parameters fixed. That is,

$$L(\beta, \sigma^2, \theta|y) = P(y|\beta, \sigma^2, \theta)$$

where $L(.)$ is the likelihood, $P(.)$ is a probability density, and y is the entire N-dimensional response vector, $N = \sum\limits_{i=1}^{M} n_i$. Because the non-observable random effects bi, $i = 1, ..., M$ are part of the model, we must integrate the conditional density of the data given the random effects with respect to the marginal density of the random effects to obtain the marginal density for the data. We can use the independence of the b_i and the ϵ_i to express the latter as,

$$L(\beta, \sigma^2, \theta|y) = \prod_{i=1}^{M} P(y_i|\beta, \sigma^2, \theta) = \prod_{i=1}^{M} \int P(y_i|\beta, \sigma^2, \theta, b_i) P(b_i|\sigma^2, \theta) db_i$$

Descriptions and comparisons of the various estimation methods used for LME models can be found, for example, in [6] and [7].

4 Illustrative Application

We here analyzed the Landsat ETM+ scene data set which comes in [10]. The ETM+ images had been previously ortho-rectified by Iranian National Cartography Center (NCC) with a high geometric precision. The geometric precision of images was also verified using road vector layer and field collected GPS control points. The aim of this study is to evaluate relationships between forest characteristics as dependent and ETM+ bands and vegetation indices as independent variables. Stepwise regression analysis selects a subset of independent variables that explains most of the variability in the dependent variable.

Independent variables of the final model were selected based on a combination of both their individual contribution to the model, adjusted coefficient of determination and residual mean square. We analyzed this data based on linear mixed effect model. Linear and linear mixed effect models were built to investigated the effect of the forest characteristics on stand volume and tree density.

4.1 Applied Methods

In the considered data set, the special inventory design of the SNFI, in which each plot is composed of four circular sub-plots with the same centre and different radial and minimum diameter threshold. It is determined that each tree i in each plot j has an unequal selection probability. Although this unequal selection probability scheme was mainly chosen for cost and administrative reasons, the hierarchical population structure underlying such schemes is of interest from a modeling point of view. It is usually argued that when the sample selection probabilities are related to the response variable even after conditioning on covariates of interests, the conventional estimators of the model parameters may be (asymptotically) biased. In such cases, weighted regression analysis with the weighting factor equal to the inverse of the selection probability leads to unbiased estimations. Linear or nonlinear regression (tree regression) can then be applied to this expanded data. By applying the same regression model to the SNFI data-type, and using weighted regression, with fixed p_{ij} as the weighting factor, the same parameter estimates are obtained. The same results are obtained when all the weighting factors are multiplied by a constant, so that they can be calculated on the basis of any per unit area, all providing unbiased estimates of the real population relationship parameters. The problem here is to choose the correct weighting factor, because the real population is unknown. In fact four statistical criteria obtained from the residuals were examined: the root mean square error (RMSE); the coefficient of determination (R2); the mean bias (E); and Schwarz's Bayesian Information Criteria (BIC; Schwarz, 1978) under squared error loss as follows [10, 11];

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} RF\exp_{ij} \times (Y_{ij}-\hat{Y}_{ij})^2}{N-p}}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N} RF\exp_{ij} \times (Y_{ij} - \hat{Y}_{ij})^2}{\sum_{i=1}^{N} RF\exp_{ij} \times (Y_{ij} - \overline{Y})^2}$$

$$\overline{E} = \frac{\sum_{i=1}^{N} RF\exp_{ij} \times (Y_{ij} - \hat{Y}_{ij})^2}{N-p}$$

$$BIC = N \times \ln \left(\frac{\sum_{i=1}^{N} RF\exp_{ij} \times (Y_{ij} - \hat{Y}_{ij})^2}{N-p} \right) + p\ln(N)$$

where Y_{ij}, $\hat{Y}_{ij}$ and $\overline{Y} = RF\exp_{ij} Y_{ij}/N$ are the measured, estimated and average values of the dependent variable respectively, and p is the number of models parameters.

Another grouping regression analysis is tree regression method. Using recursive partitioning, a sample is repeatedly subdivided on the basis of predictor variables into groups that are as internally homogeneous as possible, in terms of the outcome. For a continuous outcome such as our example, this means minimizing the within-group variance, while maximizing the differences between groups; for a categorical outcome, it means finding groups that are composed of one outcome category to the greatest extent possible. The results of Landsat ETM+ scene data set are summarized in Table 1.

Table 1: Goodness of fit statistics for the considered mixed effects models. The values in are calculated with the fixed part of the parameter estimates only.

Equation	R^2	RMSE	$\overline{E}$	BIC
(1)	0.895	1.911	-0.1001	33892
(2)	0.8847	2.003	0.0010	36326
(3)	0.8902	1.954	0.0396	35083
(4)	0.8874	1.980	0.0578	35724

In addition, we examining the scatter plot (Figure 1). As may well be the case, since the observations are taken in different species on the same study area, the Figure 1 incorporates a flexible mechanism for specifying correlation structures, and supplies constructor functions for several such structures. Most of these correlation structures, however, are appropriate only for equally spaced observations. An exception is each plot, the correlation between the appeared

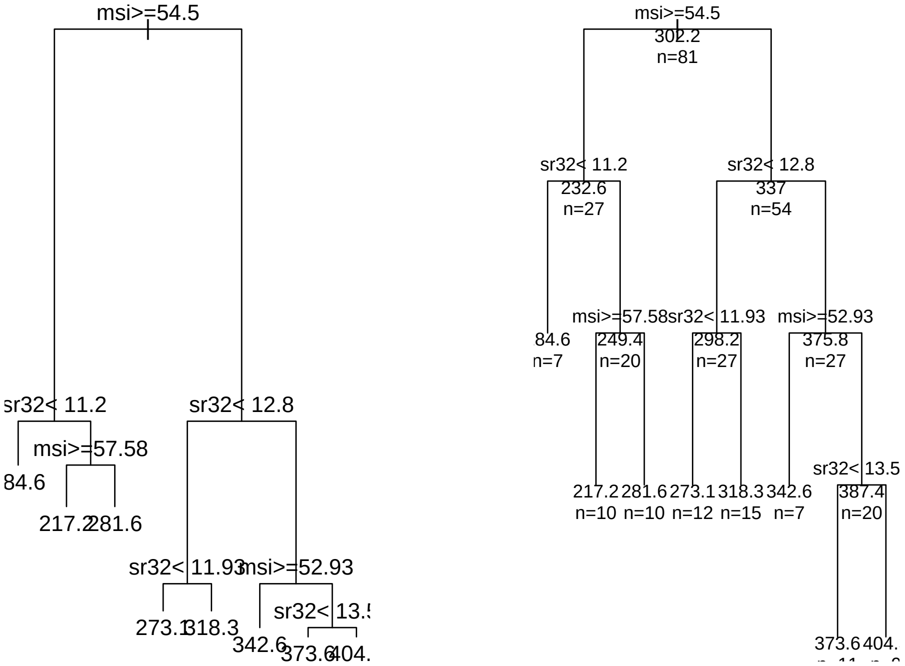

Figure 1: Variables actually used in tree construction

tress regression trees. This appears a reasonable specification in the current context, where there are at most $n_{ij} = 5, ...11$ species per plot. The model bias did not follow the same patterns in all considered models. As the tested samples are biased, they tend to increase bias in some way. The only method that maintains bias at a low level is the random selection (Figure 2). However, this procedure makes it necessary to repeat the selections, in order to avoid large errors if an inadequate tree is selected, and at least five trees should be randomly selected for the same level of accuracy as achieved by selection of the two smallest trees in the plot.

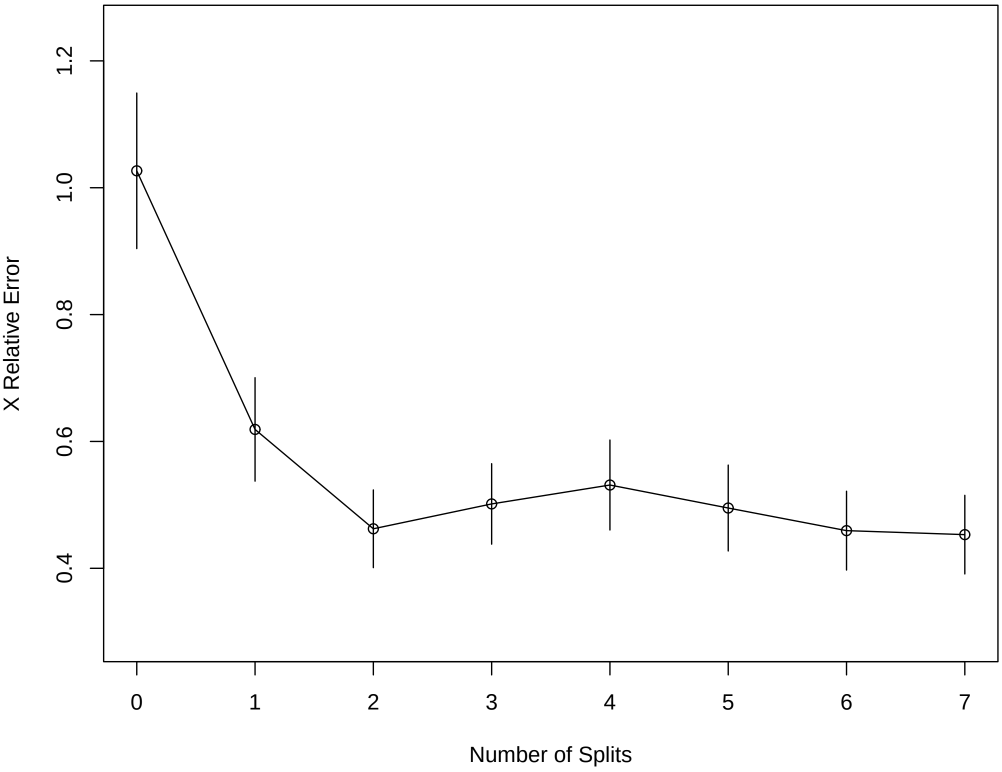

Figure 2: Defaults to one for all variables. These are scaling to be applied when considering splits, so the improvement on splitting on a variable is divided by its cost in deciding which split to choose.

5 Conclusion

A mixed model is a statistical model containing both fixed effects and random effects. These models are useful in a wide variety of disciplines. They are particularly useful in settings where repeated measurements are made on the same statistical units, or where measurements are made on clusters of related statistical units.

In this paper we compare three methods of mixed model. In the analysis of considered data set [10], from models tested in this study, the equations that showed the best fit to the dataset included dominant diameter and dominant height in their formulation. In general, the inclusion of stand variables in all models reduces bias and increases precision.

In the present study, models that included the species as a nested mixed effect provided more accurate results than those including this or are done without nested mixed effect models. This may be advantageous, since fewer trees need to be measured to estimate the volumes effect than to estimate mean number of species in each considered plot (unless the mean tree density is obtained as the tree density of the average species richness, which is, on the other hand, a less accurate way of obtaining the mean tree density), and this great sampling effort may limit future use of the models).

The best results for the calibration were obtained by selecting the smallest tree density in the plot. This may be attributed to the fact that the tree species of the plot was already considered as a fixed effect in the basic model (2) and, therefore, corresponding to the tree species did not provide much additional information for calibrations [10, 11].

In addition, the fact that the model is restricted to pass through the point implies that it cannot change much in this part of the species in nested relationship. In contrast, although measurement of the smallest tree density per plot provides a biased sample, the accuracy was greater than that of the fixed effects model, and even in comparison with the calibrated model with the randomly selected species. The greater the number of measurements included in the subsample, the greater the decrease in RMSE and increase in R^2 (Table 1).

Acknowledgement

This research was granted by the University of Golestan.

Appendix

To analyze the Landsat ETM+ scene data set, I used the nlme library in implementations of the R Software version R-2.15.1 by using of lme and nlme functions. The current version of the nlme library for R supports the same range of graphics presentations as does the S-PLUS version. This function may cover a large number of practical applications of mixed-effects models, but does not include generalized linear mixed-effects models.

For Tree regression analysis, there are two common packages in R: tree and rpart. Rpart package is easier to do that the tree. Rpart in fact implements alternative splitting functions for fitting a classification tree when interest lies in predicting an ordinal response. Some R-codes are follows;

```
model.lme¡- lme(N  Msi+Sr32+Swir+Ndvic, data = , random =   — )
summary(model.lme )
library(party)

model.rpart¡-rpart(n msi+sr32+red+swir3+ndvic,data=data, method=”anova”)
summary(model.lme ) plotcp(m)
par(mfrow=c(1,2))
rsq.rpart(m)
plot(m)
text(m)
plot(m, uniform=TRUE)
text(m, use.n=TRUE, all=TRUE, cex=.8)
par(mfrow=c(1,2))
model¡-ctree(n msi+sr32+red+swir3+ndvic,data=data)
plot(model)
library(randomForest)
model=randomForest(n msi+sr32+red+swir3+ndvic,data=data)
print(model)
importance(model)
```

References

[1] M. Davidian and D.M. Giltinan. *Nonlinear Models for Repeated Measurement Data*, Chapman Hall, London, 1995.

[2] N. A. C. Cressie, *Statistics for Spatial Data*, Wiley, New York, 1993.

[3] N. M. Laird and J. H. Ware, Random-effects models for longitudinal data. *Biometrics*, **38**, (1982), 963–974.

[4] M. J. Lindstrom and D. M. Bates, Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, **83**, (1988), 1014–1022.

[5] M. J. Lindstrom and D. M. Bates, Nonlinear mixed-effects models for repeated measures data, *Biometrics*, **46**, (1990), 673–687.

[6] E. F. Vonesh and R. L. Carter, Mixed-effects nonlinear regression for unbalanced repeated measures. *Biometrics*, **48**, (1992), 1–18.

[7] P. Adame, M. del Ro and I. Canellas, A mixed nonlinear height-diameter model for pyrenean oak (Quercus pyrenaica Willd.). *Forest Ecology and Management*, **256**, (2008), 88–98.

[8] E. F. Vonesh and V. M. Chinchilli, *Linear and Nonlinear Models for the Analysis of Repeated Measures*, Marcel Dekker, New York, 1997.

[9] R. Q. Ramos and S. G. Pantula, Estimation of nonlinear random coefficient models. *Statistics and Probability Letters*, **24**, (1995), 49–56.

[10] J. Mohammadi, S. Shataee and M. Babanezhad, Estimation of forest stand volume, tree density and biodiversity using Landsat ETM+ Data, comparison of linear and regression tree analyses. *Procedia Environmental Sciences*, **7**, (2011), 299–304.

[11] R. J. Hall, R. S. Skakun, E. J. Arsenault and B.S. Case, Modeling forest stand structure attributes using Landsat ETM+ data: Application to mapping of aboveground biomass and stand volume. *Forest Ecology and Management*, **225**, (2006), 378–390.