Conference on Applied Statistics in Agriculture 2004 - 16th Annual Conference Proceedings

AUTOMATIC MODEL SELECTION IN THE MIXED MODELS FRAMEWORK

Matthew Kramer

AUTOMATIC MODEL SELECTION IN THE MIXED MODELS FRAMEWORK

Matthew Kramer
Biometrical Consulting Service, ARS (Beltsville, MD), USDA

Abstract

Stepwise model selection is a commonly used technique in regression when there are many candidate independent variables and limited time to develop a model. This approach was adapted to the mixed models framework and gives good results, established by simulation with a known model and by application to real world data. Model selection is done using an information criterion (selected by the user). The application is primarily written in Perl. The Perl code tracks which variables are in or out of the model, calculates the information criterion, and writes and submits SAS code. Proc Mixed in SAS is used to compute the log-likelihood for a model, which is used to calculate the information criterion, which then is used to judge whether the model has improved by adding or dropping a variable, or by changing the covariance structure of the residuals. The software is currently restricted to the case where the random part of the model is assumed to be known, but how to augment the software to also select the structure for the random part of the model is discussed.

1 Introduction

Stepwise regression is a well accepted statistical method and is useful when there are many candidate independent variables but little time to develop a statistical model. While one may argue that models selected using a stepwise approach may be deficient in various ways, they can serve as both a starting point for a more in-depth model building exercise and as a reference against which other models can be compared, since a model selected in a stepwise fashion is presumably the "best" in the model space based on some criterion. With mixed model estimation available in most major statistical packages and consulting statisticians in agriculture extolling the virtues of these models, researchers often request a tool similar

to stepwise regression to identify suitable models. There are two reasons for such requests, (1) researchers believe they have only a partial understanding of these models, so that an automatic model selection procedure might avoid mistakes they could make, and (2) an automatic model selection procedure may save time if the structure of the various parts of a mixed model can be identified. This second reason also applies to consulting statisticians who have many projects and limited time, which was the motivation for developing the software described here.

This software is based on traditional partitioning of a mixed model (e.g., as described in Searle, 1971) into a fixed part (which comprises the covariates, or regression type effects, and factors), a random part (covariates and factors whose slopes or effects are sampled rather than selected prior to the experiment), and a repeated part (the covariance structure of the residuals). Model parameters are estimated using calls to SAS Proc Mixed (1999) from a Perl program, which also writes and rewrites the SAS code, explores the model space in a systematic way, and tracks improvements based on the chosen information criterion (e.g., AIC).

Others have investigated automatic model selection procedure in the mixed models framework and have made recommendations (Ngo and Brand 1997, and references therein). The only publically available code I know of for model selection in the mixed models framework is a SAS macro (also involving Proc Mixed) written by Ngo and Brand (1997), which is available online at www2.sas.com/proceedings/sugi22/STATS/PAPER284.PDF. My approach (developed independently) differs in several ways. (1) They select the fixed effects over all possible mean structures while I use a stepwise approach. Their approach is reasonable if there are few candidate independent variables. Some of the data sets I have worked with have had more than 50 candidate independent variables (the full model might not even be estimable using Proc Mixed), and all possible subsets would be time-prohibitive to run. A forward stepwise procedure runs quickly and produces good results. (2) They do not consider how Proc Mixed handles interaction terms (see below). (3) They do not reconsider the fixed part once the covariance structure is identified. (4) In my approach, a mechanism is provided for some effects (e.g., effects of treatments, blocks) to be forced into the model, regardless of whether they improve the model. (5) The application can loop over dependent variables, so that large data sets, with many dependent variables, can be analyzed more quickly. (6) Their code is written entirely in SAS and is available online, whereas mine uses both Perl code and SAS and is not yet ready for distribution. An approach similar to that of Ngo and Brand (1997) is being developed by George Fernandez, Univ. of Nevada, Reno, that will eventually become integrated into his "data mining" SAS software (personal communication).

The purpose of this paper is to demonstrate that automatic selection of the various

parts of a mixed model is feasible (even with many candidate factors and covariates), and to outline a strategy that appears to work well. In the following sections, I provide some background on mixed models and information criteria, outline the strategy used for model selection, report some simulation results, discuss real world applications, make suggestions to avoid problems, outline how to augment the method to include random effects, and make some conclusions.

2 Overview of mixed models and information criteria

Mixed models are a generalization of linear models to accomodate random effects and correlated residuals. Model estimation implementated by SAS is based on the traditional literature, e.g., Searle (1971), and therefore is somewhat limited in possible correlation structures (for the random and repeated parts). However, typical agricultural experimenters will find these sufficient.

In traditional notation used by SAS, $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $\text{var}(\mathbf{y}) = \mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$, where $\mathbf{X}$ and $\mathbf{Z}$ are the design matrices for the fixed and random effects, respectively, $\boldsymbol{\beta}$ is the vector of fixed effect parameters, $\mathbf{G}$ is the variance-covariance matrix for random effects, and $\mathbf{R}$ is the variance-covariance matrix for errors. The random effects can be factors or covariates, with the vector of random effect predictors denoted by $\hat{\boldsymbol{\gamma}}$. These solutions are shrunk towards zero, as determined by $\mathbf{G}$ and $\mathbf{R}$. While the covariance structure of $\mathbf{G}$ is typically simple (though not necessarily so), the covariance structure of $\mathbf{R}$ is often believed to be more complex and should be modeled as appropriate for the experiment (e.g., as temporally or spatially correlated residuals). Unlike the situation for general linear models, where the only covariance parameter estimated is σ^2, Proc Mixed estimates variances and parameters characterizing the structure of both $\mathbf{G}$ and $\mathbf{R}$, so $\mathbf{V}$ may be of high dimension.

Two different log-likelihood functions can be maximized with Proc Mixed. The logarithm of the restricted maximum likelihood ($l_R(\mathbf{G}, \mathbf{R})$) adjusts for the number of parameters estimated in the fixed effects part of the model, while the logarithm of the maximum likelihood ($l(\mathbf{G}, \mathbf{R})$) does not.

$$\begin{aligned}
l(\mathbf{G}, \mathbf{R}) &= -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\mathbf{r}'\mathbf{V}^{-1}\mathbf{r} - \frac{n}{2}\log(2\pi), \\
l_R(\mathbf{G}, \mathbf{R}) &= -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| - \frac{1}{2}\mathbf{r}'\mathbf{V}^{-1}\mathbf{r} - \frac{n-p}{2}\log(2\pi), \\
\text{where } \mathbf{r} &= \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{y} \text{ for both likelihoods,} \\
\text{and } p &= \text{rank}(\mathbf{X}).
\end{aligned}$$

Note that $\hat{\mathbf{V}}$ is used for estimation of $\boldsymbol{\beta}$ as $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}$, so that whatever is specified in the random and repeated statements of SAS will affect $\hat{\boldsymbol{\beta}}$.

The function $-2 \times$ (log-likelihood) + (a bias correction ["penalty"] term) is the general form of an information criterion. The bias correction term is based on the number of parameters in the model and possibly other quantities. Many information criteria have been developed starting with AIC ("Another" or Akaike's Information Criterion—see Burnham and Anderson (1998) for a history and explanation of information criteria). "Better" models have lower IC's. Information criteria can be calculated for mixed models, and thus can be used for model selection. Ngo and Brand (1997) used AIC with their method.

A major advantage of using an IC is that, with a single calculated number for a given model, calculations for comparing models are easy to program. Note that the characteristics of "better" depend on the IC used, which can change how the models are ordered. Because IC's were developed as a way of determining a directed distance between the true model and an alternative model, their optimal use in stepwise model selection is not yet settled. Models selected using an IC, however, generally minimize out-of-sample prediction error (for example, one step-ahead forecast error for time series models), which is an often a desirable property.

3 Strategy

The approach described here was first developed under the assumption that both the random and repeated structures of the model are known, with only the fixed part of the model to be determined. This approach was later modified to also determine the most appropriate repeated structure. As I have not yet had a need to also search over the random part of the model, this aspect of model selection has not been researched, although I do describe in Section 7 two strategies one might employ. Typically, a dataset has many dependent variables, and an appropriate model is needed for each.

The outline for the case with both the random and repeated parts known is the following.

1. Options are chosen (e.g., which IC to use).

2. Perl code creates SAS statements that prepare the data for Proc Mixed. If only an interaction term is specified (e.g., `y = A*B`), Proc Mixed also adds in the main effects (this can be seen by following the degrees of freedom). Thus, if one wants to allow only the interaction terms in the model, they must be created in the data step,

outside of Proc Mixed. One might want such a term if, for example, one believes two treatments to have different slopes but share the same intercept.

3. Perl code creates SAS statements that define the general model (candidate independent variables, random and repeated structures, variables forced into the model) for Proc Mixed.

4. For each candidate independent variable, Perl code is used to write SAS code (with `model y = x`, where `x` is a candidate independent variable), submits it, and parses the SAS output file.

5. Perl code is used to calculate the information criterion (the IC calculated by Proc Mixed is not usable since it includes only the number of parameters for the repeated structure in the bias correction term) and compares that with the lowest IC from previous models. If the IC using the current independent variable is lower than the previous lowest IC, the current independent variable is favored. In this way, the single independent variable producing the lowest IC is selected.

6. The candidate independent variables (now one fewer) are again checked, one by one, to determine if the IC is lowered for a model with two independent variables (one of these is the previously selected independent variable). This procedure continues so that models are systematically searched in a forward selecting manner. Independent variables are also tested to determine if they can be dropped from the model once new variables enter.

7. When no additional variables can be added or dropped, terms in the model producing the lowest IC are written as output.

8. The steps are repeated for all dependent variables, and a model is selected for each.

To be more useful, the code was modified to also find the appropriate repeated structure (still assuming the random part is known) as follows.

- First find the best set of variables for the fixed part with only the random structure given (no repeated part).

- When the fixed part is identified, identify the repeated structure (scope of the repeated part restricted to appropriate structures) by trying each in turn and keeping the one that produces the lowest IC.

- Using the identified repeated structure, determine whether the fixed part needs to be modified by adding or deleting variables. In general, few (or no) changes have been necessary for the fixed part after the repeated structure was identified.

Figure 1 is a flow chart (starting with step (4) above and for one dependent variable) illustrating the strategy used to find the fixed part of the model and the appropriate repeated structure.

4 Simulation results

To investigate this method with a known true model, data sets were simulated with 320 observations generated from a model with fixed effects of various magnitudes (from none to strong) and a known error correlation structure (AR(1), $\phi = 0.5, \sigma^2 = 4$), envisioned as a repeated measures design with 80 independent subjects, with each measured four times. The known random structure was a block (subject) effect. Results from using the approach on 1000 simulated data sets for the fixed effects are given in Table 1, with the proportion included in the final model for each effect. Effects were selected using restricted maximum likelihood and AIC corrected, with an additional penalty of 2.0 (AIC_c had to decrease by at least 2.0 before a term could be added). Various covariance structures were also tested as outlined in the previous section, with results given in Table 2.

In general, the proportion of simulated data sets that included each effect in the final model mirrored the magnitude of the true effect. Relatively unimportant effects were rarely included and important effects were usually included (given the relatively large value of σ^2, the system was sufficiently noisy that even a large effect would not be expected to always be identified as important). The only exception was the A×B interaction (A and B were both fixed effect factors, i.e., "main effects"), which was included in 20.5% of the models even though its true effect was zero. Although I am not able to provide a completely satisfying explanation for this, I ran some simulated data sets individually and found some likely explanations. Occasionally Proc Mixed estimated the random effect to be zero. When this occurred, estimates of the other covariance parameters and standard errors of the means of the treatment combinations were affected, which could result in a "significant" interaction effect. Also, during the forward selection procedure, if the B main effect was not included (the typical outcome), the A×B interaction sometimes was. The A×B interaction would likely have been included in the final model far less often had the interaction term not been allowed into the model without accompanying main effects.

Results from identifying the best repeated structure for these simulated data sets are now presented (recall that the true repeated structure was an AR(1)). The AR(1)

repeated structure was chosen most often (73.6%), with Toeplitz(2) the only other structure routinely chosen. These covariance structures are similar in that lag 1 covariances are large and covariances for large lags are small or zero. For an AR(1) parameter of 0.5, a detectable difference would only occur at small lags, such as lag 2 and lag 3 (lag 3 was the largest lag for which a covariance could be estimated in these simulations); a Toeplitz(2) would be zero at both of these lags and an AR(1) would be 0.25 and 0.125, respectively. Apparently the sample size was insufficient for these two covariance structures to be reliably distinguished.

5 Real world examples

This method was originally developed for use with data collected in a cross-over study on the effect of a diet additive on various blood components (Judd et al., 2002), and has been used when modeling data from diverse subjects at the Beltsville ARS station when blocking and temporal correlations precluded the usual stepwise regression procedures (e.g., Carroll and Kramer, 2003; Phillips et al., 2004).

Experience suggests that the larger the data set (a large data set would have 20 or more observations for each estimated parameter), the more conservative should be the IC, otherwise "superfluous" interaction terms may be added to the model (terms that only slightly decrease the IC, whose effects are difficult to detect visually from graphs, and which are difficult to interpret). In fact, it was useful to subtract a small additional penalty term from the IC when determining whether a candidate independent variable should enter the model (e.g., the IC must be lowered by at least two for a variable to enter). The order the independent variables entered was usually a good indicator of their importance, as in stepwise regression. The procedure does not guarantee that all the variables in the final model are significant (based on F-tests), just that the final combination of fixed effect variables yields the lowest IC. However, when using a conservative IC (such as BIC—Bayesian Information Criterion (see Burnham and Anderson [1998])) and the additional penalty, all the fixed effects were usually statistically significant.

6 Caveats

The most important step in preprocessing the data is to remove records that have missing values for any candidate independent variable, because IC's are not comparable for different data sets. Adding an independent variable with missing observations forces Proc Mixed to drop the records with the missing independent variables, which changes the total number of observations and thus the IC.

Centering and scaling the variables (subtracting the mean and dividing by the standard deviation) also seems to greatly improve the selection process, especially if polynomial terms are candidate independent variables, because this removes much of the co-linearity between columns of the $\mathbf{X}$ matrix. Additionally, it keeps the magnitude of covariance parameters similar. Some parameters, such AR(1), are restricted to the range (-1, 1), whereas the variance of a random effect has no upper bound. This standardization helps the algorithm find a solution more quickly and with fewer convergence problems.

Outliers can have a disproportionately strong effect on which model is selected, especially for including unnecessary interaction terms, since an outlier may make one cell (or region of a covariate) appear to be behaving differently from the others. Following model selection, it is worth performing some kind of influence analysis on the observations, for example dropping each observation in turn, re-estimating the model, and then checking for large changes in the F-values of the fixed effects. If one or more observations appear to be outliers, they should be dropped and the model selection redone.

Often, a researcher does not have a clear reason *a priori* to select one residual covariance structure over another. In such a case, letting the IC serve as a guide to determine the most appropriate covariance structure seems reasonable. The choice of the "best" covariance structure does not seem to be as important as the use of *some* covariance structure versus *no* covariance structure (the latter assumes the residual covariance structure is $\sigma^2\mathbf{I}$).

The restricted maximum likelihood function was always used (except in the very early stages of development—differences between using ML and REML were not apparent) so I have no recommendations based on experience on the choice of likelihood function. Ngo and Brand (1997), based on work of others, suggested that ML be used to determine the fixed part and REML be used subsequently.

7 Random part not known

I present two strategies for selecting the random part of the model. If there has been a constraint on randomization (e.g., blocking), then the effect capturing the constraint should always be in the model (unless this leads to estimation problems). However, one may want to determine whether interactions between random and fixed effects (which are considered as random effects) are necessary. A separate issue is to determine the correlation structure of $\mathbf{G}$. For that, a strategy similar to the one used for the covariance structure of the residuals should work.

One strategy to determine which random effects to include in the model would be to pretend the possible random structures are fixed and let the automatic modeling procedure

decide whether to retain or drop terms. The retained terms can then be included in a random statement and the model selection procedure continued. This approach is reasonable if shrinkage of the random effects towards zero would be slight. In this case, considering random variables as fixed would not greatly change their point estimates nor impact the model in other ways. This approach has the benefit of less computing time since fixed effects are estimated much more quickly than random effects. Additionally, there will be fewer problems due to non-convergence or zero estimates for random effects. The drawback with this approach is that neither fixed nor random effects will be properly tested.

A second strategy would be to loop through the candidate random effects, using an IC to determine their effect on the model in the same way fixed effects are selected, then iterate between adjusting the fixed and random parts of the models until no further changes appear necessary. Perhaps a reasonable compromise between speed (first strategy) and this second strategy is to obtain an initial estimate of likely important random effects using the first strategy, and refine the choice of both fixed and random variables with the second strategy.

Estimating a large number of models with complicated covariance structures should be avoided because that is both time consuming and is the situation where one encounters convergence problems. Once one starts trying to estimate models where, for example, parm statements are necessary for the model to find a reasonable solution, an automatic modeling algorithm is unlikely to be beneficial. However, allowing negative variances may be reasonable because it may allow the algorithm to converge (and the automatic model selection to proceed). At a later stage of model development, after examining model estimates and other information in the output of a "final" model from an automatic procedure, the user can further refine the model. This refinement might consist of adding lower bounds on variance components, using the parms statement to see if the global optimum was attained, checking for outliers, dropping terms that seem unnecessarily complex or difficult to interpret, etc.

8 Summary and Conclusions

A method for implementing an automatic model selection procedure in the mixed models framework has been outlined. The method was coded in Perl and SAS and works by iterating between the fixed and repeated parts of the model when the random part is assumed to be known. The method gave good results based on simulations from a known model and on application to real data. While the method is currently restricted to the case with the random part of the model known, ways the method might be extended to

also automatically select the random part were discussed.

The method re-examines and modifies the fixed part of the model after establishing the best covariance structure for the residuals. In general, with that strategy only small modifications to the fixed part occurred. Thus, with no random variables, one could obtain a reasonable model by (1) using ordinary stepwise regression (ignoring possible correlations among residuals) to obtain the variables to use for the fixed part of the model, then by (2) trying potential residual covariance structures with any package with a mixed models procedure, using the variables found important by the stepwise procedure for the fixed part, and an information criterion to decide which residual covariance structure is best.

Clearly, not every experiment designed in the mixed models framework will yield results amenable to automatic model selection. However, many experiments will, perhaps the majority, and software such as this should greatly reduce the amount of time both researchers and their consulting statisticans need to spend in the model building process.

9 Acknowledgments

I thank Mary Camp and Bryan Vinyard for critically reviewing the manuscript and offering helpful suggestions, and for discussions about the software while it was being developed. I thank an anonymous reviewer for many editing improvements and for suggesting I determine why the interaction term (incorrectly) became part of the final model so often for the simulated data sets.

10 References

Burnham, K.P. and D.R. Anderson. 1998. Model selection and inference: A practical information-theoretic approach. Springer-Verlag, NY.

Carroll, J.F. and M. Kramer. 2003. Winter activity of *Ixodes scapularis* (Acari: Ixodidae) and the operation of deer-targeted tick control devices in Maryland. J. Medical Entomology 40, 238–244.

Judd, J.T., D.J. Baer, S.C. Chen, B.A. Clevidence, R.A. Muesing, M. Kramer, and G.W. Meijer. 2002. Plant sterol esters lower plasma lipids and most carotenoids in mildly hypercholesterolemic adults. Lipids 37, 33–42.

Ngo, L. and R. Brand. 1997. Model selection in linear mixed effects models using SAS Proc Mixed. Technical Proceedings. SAS Users Group International, 22.

Phillips, P.L., J.B. Welch, and M. Kramer. 2004. Seasonal and spatial distributions of adult screwworms (Diptera: Calliphoridae) in the Panama Canal area, Republic of

Panama. J. Medical Entomology 41, 121–129.

SAS Institute Inc. 1999. SAS/STAT User's Guide, Version 8, SAS Institute Inc., Cary, NC.

Searle, S.R. 1971. Linear Models. John Wiley & Sons, N.Y.

11 Tables

Table 1. Simulation results for fixed effects giving the proportion of 1000 simulationed data sets that included each effect in the final model.

Effect name	Factor/Covariate	True effect/slope	Proportion included
A (2 levels)	F	2.0	0.836
B (2 levels)	F	0.5	0.033
A × B	F	0.0	0.205
time (4 levels)	F	0.0	0.001
b1	C	0.1	0.033
b2	C	0.2	0.069
b3	C	0.3	0.162
b4	C	0.4	0.268
b5	C	0.5	0.390
b6	C	0.6	0.572
b7	C	0.7	0.697
b8	C	0.8	0.830
b9	C	0.9	0.900
b10	C	1.0	0.954
b11	C	1.1	0.973

Table 2. Simulation results for repeated covariance structures giving the proportion of simulations that each covariance structure was chosen (cs = compound symmetry, vc = variance components).

SAS name	Covariance structure	Proportion chosen
ar(1)	$\sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$	0.736
cs	$\begin{bmatrix} \sigma^2+\sigma_1 & \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma^2+\sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma^2+\sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_1 & \sigma^2+\sigma_1 \end{bmatrix}$	0.000
toep(2)	$\sigma^2 \begin{bmatrix} 1 & \rho & 0 & 0 \\ \rho & 1 & \rho & 0 \\ 0 & \rho & 1 & \rho \\ 0 & 0 & \rho & 1 \end{bmatrix}$	0.247
arma(1,1)	$\sigma^2 \begin{bmatrix} 1 & \gamma & \gamma\rho & \gamma\rho^2 \\ \gamma & 1 & \gamma & \gamma\rho \\ \gamma\rho & \gamma & 1 & \gamma \\ \gamma\rho^2 & \gamma\rho & \gamma & 1 \end{bmatrix}$	0.013
vc	$\sigma^2\mathbf{I}$	0.004

12 Figures

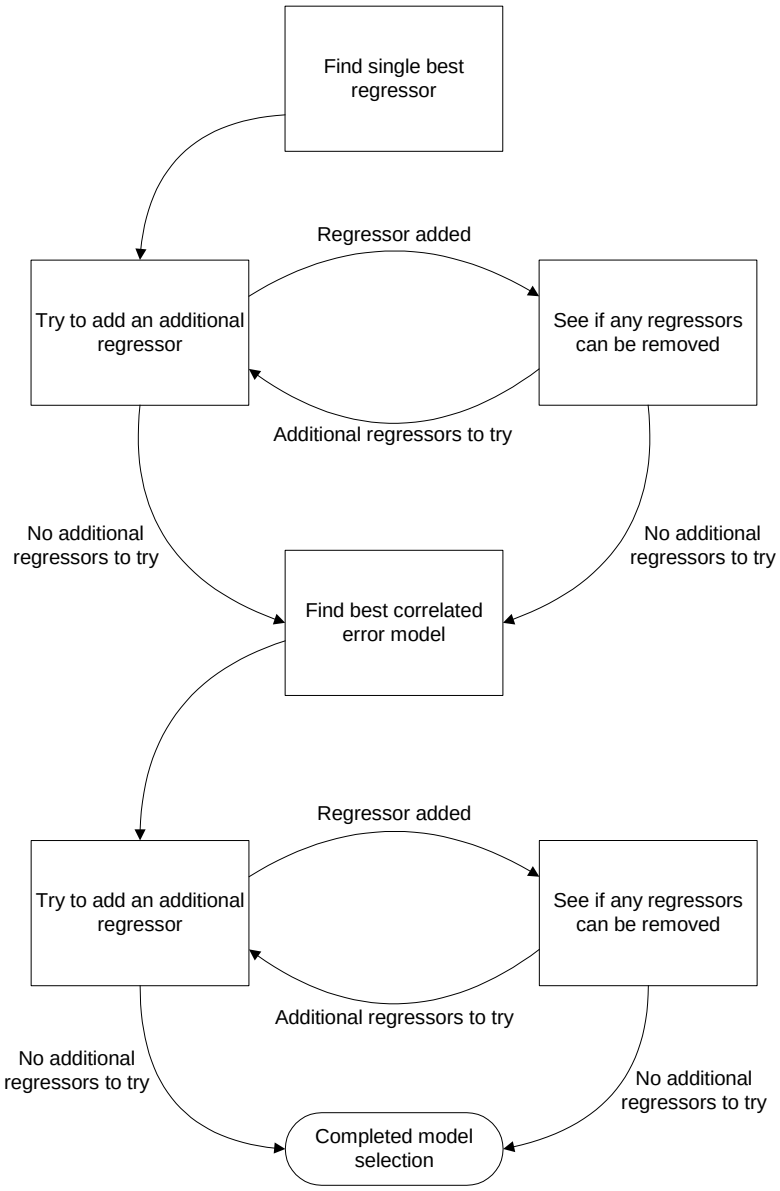

Figure 1. Flow diagram for the automatic model selection strategy described in the text for mixed models with known random effects.