SciMedCentral

Annals of Biometrics & Biostatistics

Review Article

A Variable Selection Algorithm Incorporating Investigator Preference and Missing Proportions for Linear Mixed Model

Abu Minhajuddin[1] and Hrishikesh Chakraborty[2]*

[1]*Department of Clinical Sciences, University of Texas Southwestern Medical Center, USA*
[2]*Department of Epidemiology and Biostatistics, University of South Carolina, USA*

***Corresponding author**
Hrishikesh Chakraborty, Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, SC, USA, Tel: 803-777-3170; Email: rishic@mailbox.sc.edu

Submitted: 24 October 2013

Accepted: 12 November 2013

Published: 14 November 2013

Keywords

- Backward elimination
- Regression analysis
- Linear model
- Linear mixed model
- Variable selection
- Investigator preference
- Missing proportions

Abstract

Variable selection in the context of a linear model or a linear mixed model is a fundamental but often contentious part in applied statistical model building. However, very little on the topic is available in statistical literature. In the current article, we propose a new algorithm for variable selection in the context of a linear mixed model that considers investigator preference and data availability along with other statistical considerations. The performance of the new algorithm is contrasted with the available automated variable selection using backward elimination via a real data set.

INTRODUCTION

Variable selection in the context of a linear model or a linear mixed model is a fundamental but often contentious part in the applied statistical model building. In most applied statistical research, the investigators often face the dilemma of selecting a small number of most "important" characteristics to be included in the final linear regression or logistic regression model. However, differences in characteristics selected have a direct impact on the results of the study and thus have practical consequences on how the results of the study would be interpreted and utilized. Studies in the area of social and behavioral sciences often collect a large number of relevant information on each subject. Longitudinal studies collect repeated data on these variables at different time intervals. However, because of obvious constraints such as money and time, the sample size utilized in these studies may not always be large enough. At the analysis stage, the investigators then have the difficult task of selecting a smaller sub-sample of variables that are available using some statistical criteria. The most common statistical methods for this purpose are known as forward selection, backward selection, and stepwise selection [1]. Also available are empirical Bayes' method [2], Lasso method [3], and Gibb's sampling method [4] to name a few. These methods use strictly statistical criteria such as AIC [5], BIC [6], or C_p [7] to identify the best possible set of predictors from among a much larger set. While such statistical criteria assure objectivity in model selection, they lack subjective input from the experts in the field. The subject area experts, from their experience in the area of study and focus of the analyses, may provide useful guidance in model building and thus make the end results more practically oriented towards the goals of the study. Statistical model building is a balanced combination of art and science, where the statistical criteria provide the science component and the subjective input from the experts serves as the art.

Missing data is essentially a part of any applied statistical study and a nuisance, but also can be a severe constraint on statistical analyses. Even given best efforts, it is not always possible to avoid missing data, especially in a longitudinal study. There are various types of missing data: missing completely at random (MCAR), missing at random (MAR), and non-ignorable missing (NINR) [8]. Various authors discuss methods of data analyses that can be employed when some data are missing. See [9] for a review of available statistical techniques and their properties. A comparison of available statistical methods for incomplete data regression models and various software implementations of such techniques is provided in [10].

One common feature of the variable selection methods mentioned above is that all of them assume a complete data set. In most cases this results in a complete case analysis where the cases with missing variables are deleted or the missing data are imputed using one of several imputation methods. Imputation methods, though attractive in some specific situations, are complicated, subject to additional assumptions about the data

generation process, and difficult to implement using standard statistical software. As a result, oftentimes, researchers use only the complete data cases.

In this article, we discuss a new algorithm for variable selection in the context of a linear mixed model that considers investigator preference and data availability along with other statistical considerations in statistical model building. The focus of the new algorithm is threefold: 1) to maximize the use of available data, 2) to incorporate subjective investigator input, and 3) to maintain statistical objectivity by utilizing statistical decision rules. The rest of the paper is organized as follows: Section 2 describes the new algorithm, Section 3 discusses an implementation using a real dataset, and Section 4 contains some concluding remarks.

Weighted Backward Selection Algorithm

The weighted backward selection algorithm proposed in this article is essentially a backward selection algorithm where the initial model includes all variables of interest. This complete model is then reduced to a more parsimonious model by removing some redundant variables from the initial model. The usual backward selection model removes such redundant variables based on strict criteria of statistical significance. In the weighted backward selection algorithm other considerations are also used while deciding which one variable to drop at each step. In the present article, we will demonstrate the algorithm with two additional criteria: the amount of missing data on a variable and the investigator preference of a variable. However, one can easily incorporate other considerations in the variable selection process.

The steps of the weighted backward selection algorithm where investigator preference and amount of missing data are considered along with statistical significance are described below:

Step A: Compute the percent of missing data for each independent variable.

Step B: List the investigator rankings of independent variables. The most important variable according to the investigator gets the lowest rank and so on.

Step 1: Compute the missingness index by sorting the available variables by percent of missing observations. The variable with the lowest number of missing observations gets the lowest ranking and so on.

Step 2: Estimate the model with available independent variables and create the p-value ranking for each variable. The variable with the smallest p-value gets the lowest rank and so on.

Step 3: Create the combined ranks for the independent variables by combining the three sets of rankings. Variable with high p-value rank, high missing observation rank, and low investigator rank is ranked highest.

Step 4: Exclude the one ranked highest in the combined ranking. In case of tied combined rank, the variable with the higher p-value would be dropped.

Step 5: Repeat steps 1-4 until no other variable to exclude and/or relative change in AIC (BIC) is minimal.

Figure 1 shows the details of the algorithm graphically.

Each type of ranking could either be numerical or categorical. For example, the investigator may rank the available variables into two categories: low and high. Similarly, one can group the variables into one of the three categories with low, medium, and high percent of missing observations. These rankings can be easily replaced with numeric weights. However, to avoid an infinite number of possible weighting schemes, one should assign weights so that the sum of all assigned weights of each type is one.

An Example

We demonstrate the weighted backward selection algorithm using partial data from a clinical trial on depression [11]. Data on n = 156 patients with major depression are available for a number of repeated visits. In the example, we are using data from 5 of the visits. Along with the baseline demographic and clinical characteristics, we have data on the Hamilton Rating Scale for Depression (HRSD) [12] and Beck's Depression Inventory (BDI) [13]. For patients with a significant relationship, we also have data on Dyadic Adjustment Scale (DYS) [14]. Our objective is to explore the relationship between HRSD and BDI adjusted for other factors.

Table 1 shows the combined rankings of all variables along with the missingness index, the p-value index, and the investigator rankings of the all variables at iteration one. In this example, we used categorical rankings for all three types of information under consideration. The investigator ranked the variables into one of two categories: Low (L) and High (H) while the p-value and missingness index are grouped together into four categories: Low (L), Medium (M), Medium-High (MH), and High (H). Variables with a high percent of missing observations are grouped in the high missingness category while variables with large p-values are put in the high p-value category. Thus, the variable with high missingness, high p-value and low investigator rank is assigned the highest combined rank. If there is more than one variable in that category, then the variable with the higher p-value among them is assigned the higher rank. In our example, the variable

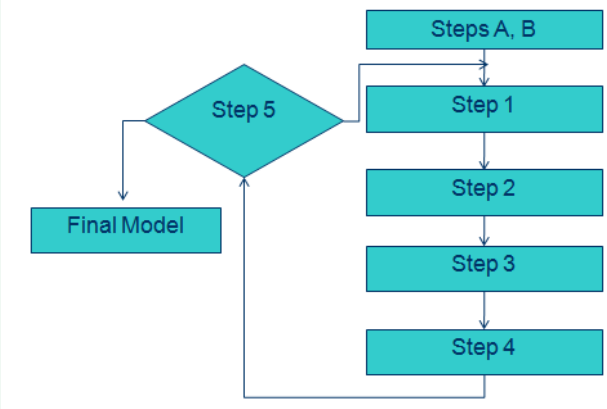

Figure 1 Weighted backward selection algorithm.

DYS and length of current episode has low investigator rank, medium-high p-value index, and high missingness index and thus assigned the highest combined rank. However, the variable DYS has the higher p-value between the two variables and thus would be dropped at the first iteration. The remaining variables would then be used to identify the next variable to be eliminated. The process will continue until either all candidate variables are eliminated or the absolute change in AIC is smaller than a pre-specified minimum.

Table 2 shows the variables eliminated at each step along with the number of missing observations and p-value of the eliminated variable, the effective sample size used, and the AIC value. It is evident that the weighted backward selection algorithm utilizes more of the available data by re-evaluating the effective sample size after each elimination. It also re-calculates the three indices as well as the combined rankings. The regular backward selection evaluates the effective sample size only at the beginning of the first iteration and thus uses a much smaller sample for all the iterations. Also, only the p-value rankings are used to decide which variable to eliminate at each round leading to a completely different set of variables to be eliminated. See Table 3 for a side-by-side comparison of the two algorithms in terms of variables eliminated and the effective sample size at each iteration.

The available variable selection algorithms such as backward, forward, or stepwise selection allow the investigator to force a variable to be included in the final model without any consideration of the statistical significance of the variable in question. The weighted backward selection algorithm, on the other hand, allows the investigator to place varying levels of importance on each variable via the weighting scheme. However, in this algorithm, the final decision to include or eliminate a variable relies on the statistical importance of the variable. In Table 1, with a low investigator ranking, DYS was eliminated in the first iteration. However, if the investigator ranking was changed from low to high, DYS would not have been eliminated until iteration eight.

Table 1: Combined rankings of variables at iteration 1 along with three types of rankings.

Variable	P-value index	Missingness Index	Investigator rank	Combined Rank
Age	L	MH	H	3
Gender	H	L	H	7
Education	M	H	L	27
Ethnicity	H	L	L	23
Employment	M	M	H	8
Paired	M	MH	H	9
RDC Primary	MH	L	L	22
RDC Endogenous	H	L	L	23
Age of Onset	L	M	L	18
Length of Current Episode	MH	H	L	30
Length of Illness	L	M	L	18
BDI	L	MH	H	3
DYS	**MH**	**H**	**L**	**30**

Table 2: Variables eliminated at each iteration.

Iteration	Dropped Variable	P-value	Number missing	N used	AIC
01	DYS	0.299	306	347	1975
02	Education	0.539	30	588	3353
03	Length of Current Episode	0.997	20	617	3505
04	Age of Onset	0.431	10	637	3610
05	Length of Illness	0.832	7	647	3669
06	RDC Primary	0.119	5	653	3691
07	RDC Endogenous	0.013	0	658	3719
08	Ethnicity	0.007	0	658	3724
09	Paired	0.011	15	658	3732
10	Employment	0.003	12	673	3829
11	Age	0.000	15	680	3874
12	Gender	0.009	0	694	4031
13	BDI	0.000	18	694	4038
14	...	...	...	712	5872

Table 3: Weighted and Regular Backward Elimination Algorithms.

Iteration	Weighted Backward Selection			Regular Backward Selection		
	Drop	N Used	AIC	Drop	N Used	AIC
1	DYS	347	1975	Gender	347	1975
2	Education	588	3353	RDC Endo	347	1977
3	LOCE	617	3505	Ethnicity	347	1978
4	AAO	637	3610	LOCE	347	1979
5	L of Illness	647	3669	RDC Primary	347	2032
6	RDC Primary	653	3691	Employment	347	2034
7	RDC Endo	658	3719	Paired	347	2062
8				Education	347	2065

CONCLUSIONS

The weighted backward elimination algorithm described here incorporates factors other than the p-value in model. While consideration of p-value alone brings objectivity to the model building process, it completely ignores other extraneous factors. In the weighted backward elimination algorithm, both missingness of observations and investigator preference is incorporated in the process. Other such factors could also be included by using more factors while computing the combined rank of variables. The proposed algorithm also maximizes the use of available data without resorting to imputing the missing data by evaluating effective sample size repeatedly. Thus the estimated model would be free of any additional assumptions required for the data imputation methods. It is also free of additional programming difficulties associated with data imputations.

Our goal in this article is to describe an automated variable selection algorithm. Thus, throughout the discussions in this article, we have assumed non-informative missing data. We have also assumed the appropriateness of the linear model as well as other simplistic assumptions required for such a model.

REFERENCES

1. Draper NR, and Smith H. Applied Regression Analysis, 2nd Ed. New York: John Wiley and Sons; 1981.
2. Yuan M, and Lin Y. Efficient empirical Bayes variable selection and estimation in linear models. J. Am. Statist. Ass. 2005; 100: 1215–1225.
3. Tibshirani R. Regression shrinkage and selection via the lasso. J. R. Statist. Soc. B. 1996; 58: 267–288.
4. George EI, and McCulloch RE. Variable selection via Gibbs sampling. J. Am. Statist. Assoc. 1993; 88: 881–889.
5. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974; 19: 716–723.

6. Schwarz G. Estimating the dimension of a model. Annals of Statistics. 1978; 6: 461-464.

7. Hocking RR. The Analysis and Selection of Variables in Linear Regression. Biometrics. 1976; 32: 1-50.

8. Little RJA, and Rubin DB. Statistical Analysis with Missing Data. 2nd ed. New York: John Wiley and Sons; 2002.

9. Molenberghs G, and Kenward MG. Missing Data in Clinical Studies. New York: John Wiley & Sons; 2007.

10. Horton NJ, Kleinman KP. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. Am Stat. 2007; 61: 79-90.

11. Jarrett RB, Kraft D, Doyle J, Foster BM, Eaves GG, Silver PC. Preventing Recurrent Depression Using Cognitive Therapy With and Without a Continuation Phase: A Randomized Clinical Trial. Arch Gen Psychiatry. 2001; 58: 381-388.

12. Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry. 1960; 23: 56-62.

13. Beck AT. Depression: Causes and Treatment. Philadelphia: University of Pennsylvania Press; 1972.

14. Daspe MÈ, Sabourin S, Péloquin K, Lussier Y, Wright J. Curvilinear associations between neuroticism and dyadic adjustment in treatment-seeking couples. J Fam Psychol. 2013; 27: 232-241.

Cite this article

Minhajuddin A, Chakraborty H (2013) A Variable Selection Algorithm Incorporating Investigator Preference and Missing Proportions for Linear Mixed Model. Ann Biom Biostat 1(2): 1010.