JOURNAL OF MULTIVARIATE ANALYSIS **58**, 27–54 (1996)
ARTICLE NO. 0038

Covariate Screening in Mixed Linear Models

A. M. RICHARDSON

University of Canberra, Canberra, Australia

AND

A. H. WELSH

The Australian National University, Canberra, Australia

We address the important practical problem of selecting covariates in mixed linear models when the covariance structure is known from the data collection process and there are a possibly large number of covariates available. In particular, we consider procedures which can be considered extensions of the analysis of deviance to mixed linear models. This approach provides an alternative to likelihood ratio test methodology which can be applied in the case that the components of variance are estimated by restricted maximum likelihood (REML), thus resolving the open question of how to proceed in this context. Moreover, it is simple to robustify and allows us to consider a wider class of procedures than those which fit into the simple likelihood ratio test framework. The key insights are that the deviance should be specified by the procedure used to estimate the fixed effects and that the estimated covariance matrix should be held fixed across different models for the fixed effects.

1. INTRODUCTION

The question of which variables to include in a statistical model is basic to much applied statistics. In this paper, we address the problem of covariate selection in mixed linear models. We adopt the viewpoint that the covariance structure is known from the data collection process and that there are a possibly large number of covariates available. The problem is to select a useful set of covariates to provide a parsimonious representation for the mean given the covariance structure.

Received May 5, 1994; revised February 1996.

AMS 1980 subject classifications: 62H15, 62H35.

Key words and phrases: analysis of deviance, hypothesis testing, likelihood ratio, mixed model, REML, robustness.

0047-259X/96 $18.00

Consider the general mixed-effects linear model of the form

$$y = X\alpha + \sum_{i=1}^{c-1} Z_i\beta_i + \varepsilon, \tag{1.1}$$

where y is an n-vector of observations; X and Z_i are known $n \times q$ and $n \times p_i$ design matrices respectively; α is a q-vector of unknown fixed effects; the β_i are p_i-vectors of unobserved random effects, $1 \leqslant i \leqslant c-1$; and ε is an n-vector of unobserved errors. The p_i levels of each random effect β_i are assumed to be independent with mean zero and variance σ_i^2; each random error is assumed to be independent with mean zero and variance σ_c^2; and $\beta_1, ..., \beta_{c-1}$ and ε are assumed to be independent. Thus

$$Ey = X\alpha$$

and

$$\mathrm{Var}(y) = V = \sigma_c^2 I_n + \sum_{i=1}^{c-1} \sigma_i^2 Z_i Z_i'.$$

We will assume for simplicity that we have adopted a parameterisation in which all the $t = q + c$ unknown parameters $\tau = (\alpha', \theta') = (\alpha', \sigma_1^2, ..., \sigma_c^2)'$ are identifiable.

The problem of screening covariates in a regression model ($c = 1$) has attracted considerable attention; see Atkinson [2] for a useful summary. The general approach is to define a measure of discrepancy D between a model and the data and then explore how changes in the model affect this measure. Akaike [1], Mallows [22] and Stone [39] suggested modifying D to $D + a_n q\hat{\sigma}_c^2$, where a_n is a deterministic sequence and $\hat{\sigma}_c^2$ is an estimate of σ_c^2, to incorporate an explicit penalty for increasing the number of covariates and then examining models which minimise $D + a_n q\hat{\sigma}_c^2$. Alternatively, we can accept a sequence of changes to a model which produce non-significant changes in D at say the 5% level. Effectively, at each step, we test the hypothesis that some covariates can be excluded using the change in discrepancy ΔD as a test statistic. After permuting the columns of X if necessary, the null hypothesis at each step can be expressed equivalently as the first $s \geqslant 1$ components of α equal zero, i.e.

$$H_0 : H_{\alpha\alpha}'\alpha = 0,$$

where the $s \times q$ matrix $H_{\alpha\alpha}' = (I_s : 0)$, I_s is an $s \times s$ identity matrix and $s \leqslant q$. Given the distribution of ΔD under H_0 and an algorithm for determining the sequence of models to examine, this screening procedure is straightforward to implement.