Computational Statistics and Data Analysis 70 (2014) 1–18

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Automating the analysis of variance of orthogonal designs

Heiko Großmann *

School of Mathematical Sciences, Queen Mary University of London, Mile End Road, London E1 4NS, UK

ARTICLE INFO

Article history:
Received 26 September 2012
Received in revised form 26 August 2013
Accepted 28 August 2013
Available online 11 September 2013

Keywords:
Analysis of variance
Orthogonal block structure
Algorithm

ABSTRACT

A new algorithm is presented which for the wide class of orthogonal designs is capable of deducing the appropriate analysis of variance from the design only. As a consequence the use of a model equation for specifying the analysis becomes dispensable. The proposed approach can simplify the analysis of complex models with iterative crossing and nesting of factors, where treatment factors have fixed and plot factors have random effects. An implementation is described and its use is illustrated with several examples.

1. Introduction

This paper is about a computer program for the analysis of variance (anova). Even before finishing the reading of the first sentence the reader may wonder if another such program is needed. In the paper I will try to answer this question in the affirmative by explaining how the proposed *AutomaticAnova* package can simplify the analysis of complex anova models with complicated blocking structures and factors having random and fixed effects in a way, I believe, that no other existing package can. Of course, this statement needs to be taken with a pinch of salt, because although being very general the theory underlying the program has its limitations and, essentially, only applies to orthogonal designs as defined by Bailey (2008). Also, in order to prevent later confusion, it seems to be appropriate to point out from the beginning that, despite being a familiar term, the name 'orthogonal design' means different things to different people and therefore being clear about the definition used in this paper is important.

The *AutomaticAnova* package originated from two sources. One was the teaching of a module *Design of Experiments* originally designed by Professor R. A. Bailey, the material of which is now available in Bailey (2008). The other was a collaboration with biologists reported in Muller et al. (2010). From the teaching it became clear that the theory had some algorithmic content which could lead to the anova being automated. The joint work with the biologists on the other hand revealed that a software implementation of Bailey's theory could tremendously reduce the time needed for providing consultancy.

At this point it seems to be appropriate to give the reader an idea about what 'automating the analysis of variance' means. In short, this phrase refers to the package's capability to *infer* an appropriate model from a design provided in the form of a spreadsheet and to carry out the analysis *without the need to specify a model equation*. From a practical point of view the fact that the user does not have to specify the model appears to be crucial, since from experience it seems that non-statisticians usually find it hard to understand and apply the operators, such as nesting and crossing, which are commonly offered by software for defining anova models. It is also worthwhile to note that the program's algorithm for deriving the model is not based on a fixed collection of predefined designs or models, but works for any orthogonal design.

In addition to performing the anova computations the program also generates Hasse diagrams, which have been recognized as a useful tool for understanding the structure of experiments by Taylor and Hilton (1981), Tjur (1984), Bergerud

* Tel.: +44 0 20 78823113; fax: +44 0 20 8981 9587.
E-mail address: h.grossmann@qmul.ac.uk.

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.08.014

(1996) and Vilizzi (2005). Lohr (1995) emphasizes applications of the diagrams for teaching and consulting purposes. Several of these authors discuss how Hasse diagrams can be used for deriving the anova table, but none considers how the diagram can be automatically extracted from the design.

The sections that follow provide a brief account of the theory on which the *AutomaticAnova* package is based, present the algorithm for automating the anova and give some information on the implementation. In addition, the practical use of the package is explained and illustrated with several examples. The paper concludes with a discussion of limitations and further extensions.

2. Bailey's theory for orthogonal designs

The analysis of variance is one of the most versatile statistical techniques in common use. Although the principles on which the anova is based are well understood, there exist different perspectives on the method. In one, the anova is regarded as an instance of the general linear model and the analysis is considered from a regression point of view (e.g., Christensen, 1987). Other presentations focus on model equations and corresponding decompositions of sums of squares and degrees of freedom (e.g., Sahai and Ageel, 2000). A third approach emphasizes randomization ideas and clearly distinguishes the structure of the observational or experimental units from the structure of the treatments (Nelder, 1965a,b). Interestingly, often proponents of different anova 'schools' have difficulties understanding each other. A thorough discussion of these matters is beyond the scope of this paper, but more information can be found, for example, in the discussion papers by Speed (1987) and Gelman (2005), in Section 4.3 of Brien (1989) and in Sections 3.1–3.2 of Brien et al. (2011).

In this paper, the focus is on a version of the third approach presented in the monograph by Bailey (2008), which generalizes the seminal work of Nelder (1965a,b). The brief summary of the theory below is necessary for understanding the material in Section 3. Related ideas have been presented by Houtman and Speed (1983), Tjur (1984), Bailey (1981, 1996) and Payne and Tobias (1992). Readers who are familiar with Bailey's approach to the anova may skip the rest of this section and only use it as a reference later on.

2.1. Definitions and notation

In what follows, for the most part I adopt the notation in Bailey (2008) to facilitate comparisons of the material in this paper with the more comprehensive account in Bailey's book. Where a modified notation is used this is motivated by trying to make some aspects of the theory more explicit.

The theory distinguishes the set Ω of size N which represents the observational units from the set $\mathcal{T}$ of treatments which has size t. A plot (or block) factor F is a function from Ω to a finite set of n_F levels and similarly a treatment factor G is a function from $\mathcal{T}$ to a finite set of n_G levels. For simplicity of presentation I assume that Ω consists of the integers $1, \ldots, N$ and that the n_H levels of every factor H on Ω or $\mathcal{T}$ are represented by the integers $1, \ldots, n_H$.

Plot factors reflect the inherent structure of the observational units, such as arrangements into blocks. Treatment factors, on the other hand, have their levels deliberately chosen and applied by the experimenter, usually after some suitable randomization. It is assumed that there are no interactions between factors of the two types. Bailey (2008, pp. 14 and 279–281) and Cox (1984) explain some of the inferential problems that arise if there are such interactions.

Bailey (1981, 1991) argues that, for the orthogonal designs considered here, performing the usual randomization justifies the model in which all plot factors have random effects and all treatment factors have fixed effects. Thus the algorithm for data analysis can be applied whenever the factors can be separated into 'plot' factors with random effects and 'treatment' factors with fixed effects, so long as there are no interactions between the two types.

Every plot factor F gives rise to a partition of Ω into the F-classes $F[[i]] = \{\omega \in \Omega : F(\omega) = i\}$ for $i = 1, \ldots, n_F$, and likewise a treatment factor G introduces a partition of $\mathcal{T}$ into G-classes $G[[j]] = \{a \in \mathcal{T} : G(a) = j\}$, where $j = 1, \ldots, n_G$. Thus each factor H with n_H levels can be identified with a set of n_H sets, which are the H-classes. Bailey (2008, p.169) defines the classes in a slightly different way, but both definitions lead to the same partitions.

There are two special factors each of which can be defined on Ω or $\mathcal{T}$. The universal factor U has only a single level and hence a single U-class. By contrast, the equality factor E has as many levels and E-classes as there are elements in Ω or $\mathcal{T}$. Since it should be clear from the context on which set U (or E) is defined I will use the same symbol U (or E) for the factors on Ω and $\mathcal{T}$.

By identifying each plot factor F with the partition $\{F[[i]] : i = 1, \ldots, n_F\}$, any given collection of plot factors can be partially ordered in a simple way. To this end, let $\mathcal{F}$ be a finite set of plot factors. Two factors $F, G \in \mathcal{F}$ are equivalent, denoted by $F \equiv G$, if they have the same classes. Otherwise they are called inequivalent. Moreover, F is said to be finer than G (or G to be coarser than F) if the factors are inequivalent and if for every $i \in \{1, \ldots, n_F\}$ there exists a $j \in \{1, \ldots, n_G\}$ such that $F[[i]] \subseteq G[[j]]$. This is denoted by $F \prec G$ or $G \succ F$. For inequivalent factors $F, G \in \mathcal{F}$, in words $F \prec G$ means that whenever two units in Ω have the same level of F then they also have the same level of G. Finally, a factor $F \in \mathcal{F}$ is finer than or equivalent to $G \in \mathcal{F}$, which is denoted by $F \preceq G$ (or $G \succeq F$), if $F \prec G$ or $F \equiv G$. Any set $\mathcal{F}$ of plot factors can then be partially ordered in terms of the relations $\prec$ or $\preceq$. Likewise, any set of treatment factors $\mathcal{G}$ on $\mathcal{T}$ can be partially ordered in terms of similarly defined relations $\prec$ or $\preceq$.

In addition to being able to separately define partial orders for sets of plot and treatment factors, new factors can be created from old ones by means of two binary operators $\wedge$ and $\vee$. Since the operators are defined in the same way for plot