Methodological aspects of statistical modelling: some new perspectives

Herman J. Adèr[1], Dirk J. Kuik[1], Jan B. Hoeksma [2], Gideon J. Mellenbergh [3]

[1] Faculty of Medicine, Department of Clinical Epidemiology and Biostatistics. Vrije Universiteit, Van der Boechorststraat 7, Amsterdam, The Netherlands.
[2] Faculty of Psychology and Education, Department of Developmental Psychology
[3] Faculty of Psychology, Dept. of Psychological Methods, University of Amsterdam

Abstract: In the paper, we will first derive a graphical specification of the translation process of a substantive problem into a statistical model and of the interpretations of the results in terms of the subject matter domain. The whole procedure is termed 'methodological modelling'.
To demonstrate a methodologically oriented approach to statistical modelling a new application of Edwards and Havránek's model search algorithm is presented appropriate for a model space of survival models. The algorithm allows for user intervention at each step.

Keywords: Substantive research problem, methodological modelling, survival analysis, model search, EH algorithm.

1 Introduction

It is amazing that general literature on statistical modelling is so scarce. Dobson (1983) seems to be the first book that discusses statistical modelling in a general way although it does not consider more advanced modelling techniques like structural equation modelling or graphical modelling. A recent source is Edwards (2000, Chapter 6) that provides an more general discussion of modelling.
The present paper has two aims: (a) We introduce the concept of *methodological modelling* and will give a graphical description of the translation of a research problem into a statistical model and of the interpretation of the results in terms of the subject matter domain; (b) We give a new application of Edwards and Havránek's model search algorithm to Cox regression analysis, adding the possibility of user intervention.
We will try to arrive at a more formal description of this modelling process using a graphical technique developed in Adèr (1995).

We will make a distinction between modelling in the 'context of justification' and in the 'context of discovery' which leads to different specifications. In the last instance, the advantages of splitting the procedure in a modelling and a testing phase, will be discussed. As an example of a statistical modelling technique that allows for a high level interpretation, we discuss a model strategy introduced by Edwards and Havránek.

2 Methodological modelling (M-modelling)

Definition 1 (Methodological modelling) Methodological modelling *indicates the subsequent activities of (a) translating a substantive research problem into the specification of a class of statistical models (*the model space $\mathcal{M}$*), (b) selecting the most relevant subset $\mathcal{S} \subset \mathcal{M}$ of statistical models (*statistical modelling*), and (c) Interpreting S in terms of the research problem.*

We also use the abbreviation *M-modelling* instead of 'methodological modelling', in contrast to *S-modelling* or statistical modelling. Note that the term 'statistical modelling' may sometimes be too narrow: it is intended to include also cases in which simulations or neural network training are used (see below).
In (a), the research problem is translated into the specification of a set of manageable statistical models. This can be done without actually specifying a research design and collecting data. Note that usually something is lost during this translation: usually the resulting statistical models are only valid under assumptions imposed by the statistical framework. This may restrict the generality of the answers the S-modelling phase provides and may lead to 'giving good answers to the wrong problem'.
As to (b): methodological modelling encompasses statistical modelling. Here we actually *need* data, so this activity implicitly presupposes that a research design has been specified and that data are collected on which to execute the analyses. As to S-modelling: there is a host of literature on various statistical modelling techniques: (a) Regression modelling; (b) Log-linear modelling; (c) Multilevel modelling; (d) Structural Equation Modelling; (e) Graphical modelling; (f) Modelling using the state-space model; (g) Dynamic modelling; (h) Neural network modelling. In the related literature, examples are given that also include the recommended interpretation in terms of the research domain.
To the best of our knowledge, nothing more general has been said on the third phase ((c) in our definition), possibly because this is on the one hand closely related to the specific subject matter domain where the research question arises and, on the other hand, to the specific statistical technique.

One may ask what is the place of methodological modelling in the dichotomy *confirmatory* versus *exploratory* research, or, to put it more generally: whether modelling fits into the 'context of justification' or into the

'context of discovery'. The answer is that it fits in either contexts but that the procedures are different. The corresponding relevant questions are: (a) *Do the data confirm the theoretical models that were postulated beforehand ?* (Theory-driven); (b) *Do the statistical analyses of the data suggest associations and structures in the subject matter domain?*(Data-driven). In the first case, a theoretical framework or even a theory should be available and it should be possible to produce a clear problem formulation in terms of the research domain. The methodological task is to translate this into a research design that, when data are collected, allows to test statistical models that are direct translations of subject matter considerations. The statistical analysis is then aiming at confirmation of a theoretical concept: the models that were specified beforehand are tested on the data, and we want to know if they are supported, in other words if they *fit* the data. The interpretation of the outcome of the statistical analysis allows an answer to the research problem in the form of a label or stamp like CONFIRMED or NOT CONFIRMED. This resembles *hypothesis testing* with the difference that in case we conclude that our hypothesis is not confirmed, we may develop ideas for a theory that *is* supported by the data. At that moment we switch to a data-driven approach, because the newly formulated theory is suggested by this data only.

In the exploratory case, no theoretical model in the strict sense is available although ideas about the relevance of certain notions for the problem are necessary to be able to formulate a research design and collect data accordingly. Once the data have been collected, statistical modelling proceeds in an exploratory way, which can be described as follows: we are 'poking around' in the data trying to find out interesting patterns in terms of the subject matter domain. Note that this, in turn, resembles aims and strategies used in data mining. In this case, there is always an interpretative problem: since this method is completely data-driven, we may arrive at conclusions that are only relevant for the data under scrutiny. In some cases a solution may lay in some kind of *cross-validation*, for instance by splitting the data in two parts and use one part for exploration ('learning') and the other to confirm what is found in the first part ('testing'). This clearly presupposes that there are enough data available to do the analyses on each part. The exploratory procedure is indicated in Figure 1. It may be obvious from the above description of confirmative and exploratory methodological modelling that each procedure holds parts of the alternative: if in the confirmatory case the predefined theory is not confirmed, a new theory is formulated (or the old theory is adapted) based on the findings during S-modelling and all of sudden our confirmatory proceeding becomes data-driven and exploratory. On the other hand, the procedure proposed above for exploratory M-modelling becomes confirmatory if we follow a cross-validation strategy. And, if we consider the scientific process in a research field, both kinds of modelling alternate: exploratory analyses are done first and, based on the results, other researchers try to confirm or falsify the findings by doing new experiments.

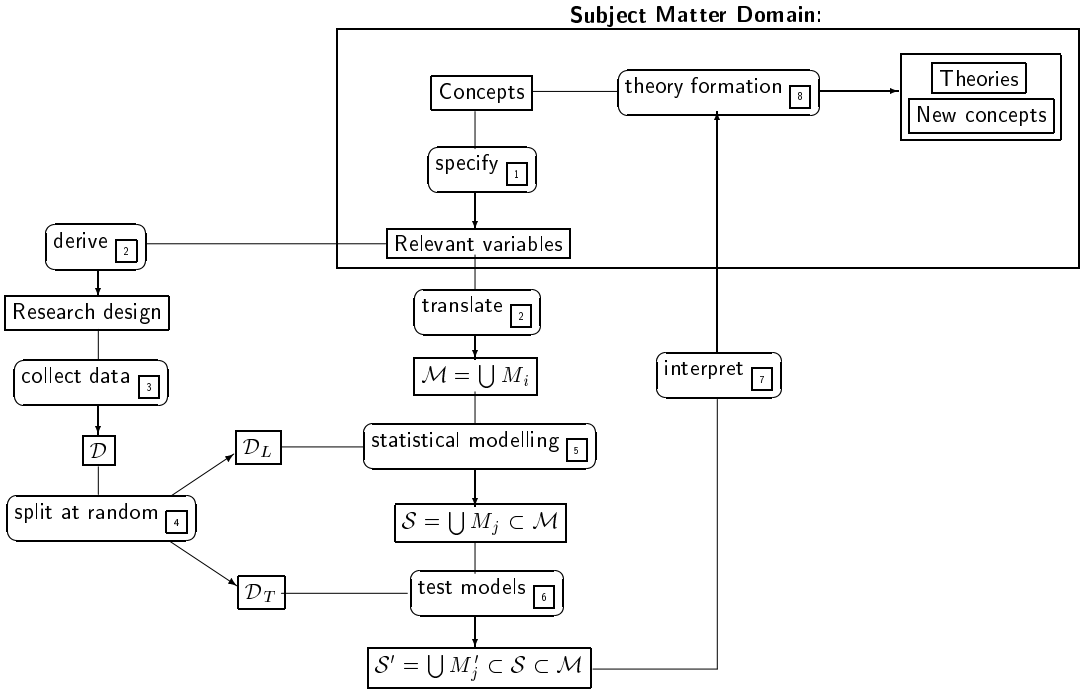

FIGURE 1. Methodological modelling in the context of discovery. $\mathcal{M}$: class of models; $\mathcal{S}$: accepted models; $\mathcal{S}'$: confirmed models from $\mathcal{S}$. $\mathcal{D}$: data, $\mathcal{D}_L$: 'learning' data set; $\mathcal{D}_T$: 'test' data set.

We now first restrict ourselves to the large class of regression models. They all have in common that we can discern two classes of variables: *dependent* and *independent* variables. The information needed for the assignment of variables to these classes can only be derived from the subject matter domain. This also applies for other information like whether a variable is focal or only a 'confounder', a disturbing variable, for which we want to correct our models.
We will further concentrate on the problem of finding a parsimonious set of independent variables that optimally predict the dependent. We will focus on Cox regression analysis, although the proposed procedure will be completely applicable to other forms of regression analysis.

3 Example data set

The example is taken from the study by Van Bokhorst-de van der Schueren et al. (2000) in which forty-nine severely malnourished head and neck cancer patients undergoing ablative and reconstructive surgery were followed prospectively and their perioperative immune parameters were related to long-term survival. In the study it was found that preoperative human leukocyte antigen-DR (HLA-DR) expression on monocytes was different between patients who survived and patients who died. Proportional hazards identified an HLA-DR expression on monocytes below the cutoff points (mean fluorescence index < 15, peak channel index < 9) to be of significant influence on survival. The authors draw the conclusion that in addition to known prognostic parameters, (with other parameters) the immune parameter HLA-DR expression on monocytes may carry prognostic value in cancer patients.
The example is used here only to demonstrate the merits of the EH procedure. We analyzed models that used six of the most promising predictors but many more variables were used in the study.

4 Model search methods

By choosing one of the many stepwise approaches, most of the variable selection process is left to the software itself, following an algorithm that is exclusively defined in statistical terms and in which there is no room for subject matter considerations. To circumvent this, some authors advocate a more content-directed iterative stepwise procedure which allows the user at each step of the iteration process to indicate to include or exclude the term that has been brought up by the software.
In Edwards's book, the so-called *EH-algorithm* is discussed. The method allows to search a large model space in an efficient way. This method is based on assumptions on this model space rather than on individual models. It allows to find, instead of only one optimal model, a whole series of

models that all have optimal properties. In the regression case, by properly sorting the resulting models, we can get information about the association of the individual variables and the dependent variable. Furthermore, collinearity does no longer influence the results in the way it occurs in stepwise regression analysis. The EH procedure can be executed using the program `MIM` (an acronym for **M**ixed **I**nteraction **M**odelling) program developed by Edwards (2000). Alternatively, we developed 'scenario' for the parallel empty shell `PROTOSHELL` that allows to search a space of logistic regression models. In the next section we will report on the development of a similar application for Cox regression analysis in which the EH algorithm is used but in which user intervention is allowed at each step.
We give a sketch of the requirements and assumptions underlying Edwards's algorithm: (a) A distance measure between models must be defined. (b) For this distance a threshold must be defined beforehand that serves to discern between accepted and rejected models. (c) A hierarchical ordening is defined on the models: model $M_A \supset M_B$ when model M_B has all the terms of M_A but M_A has more terms. M_B is said to be more *parsimonious* than M_A. (d) The assumption of *coherence* should hold, i. e. it should be reasonable to assume that if model M_A is less parsimonious than M_B and M_B is rejected, that it follows that M_A is rejected. Similarly, if $M_{A'}$ is more parsimonious than $M_{B'}$ and $M_{A'}$ is accepted, $M_{B'}$ is also accepted. (e) The procedure starts from either the empty model $M_\emptyset$ which is assumed to be rejected or the full model which is declared accepted and via the above rules tries to find the most parsimonious accepted models.

5 Results from the example data set

Variables were selected from the many variables in the data set which showed most influence in forward and backward stepwise Cox regression analysis. The following variables were used: (a) Main location of the tumor (**MAL**), (b) Swallowing (**SWA**), (c) Hospitalization duration (**HDU**), (d) Major Complications (**MAC**), (e) Minor Complications and, (f) Antigen-DR at followup (**ANF**). Table 1 summarizes the output of the EH Cox regression analysis.
χ^2 change was used as a distance between models. The threshold χ^2 value for 1 degree of freedom at $\alpha = 0.05$ is 3.841. Note however that for the categorical variables like MAL the number of degrees of freedom is larger than one. The χ^2 changes of the variables versus the empty model were: MAL: 14.3, SWA: 34.3, HDU: 6.1, MAC: 4.2, MIC: 2.2 and ANF: 3.0). Thus, MIC and ANF are not statistically different from a rejected model. By user decision they were included for testing in higher order models. Another user decision was taken in the case of model (7): χ^2 was 3.561 and it was left in the models to be tested.
The procedure was run both upwards (starting from the (rejected) empty model) resulting in models 1-10 and downwards (from the (accepted) saturated model) to yield models 11-14.

	MAL	SWA	HDU	MAC	MIC	ANF	U/A/R
One independent							
1	×						
2		×					
3			×				
4				×			
Two independents							
5	×	×					R
6		×			×		
7			×		×		UR
8			×	×			
Three independents							
9		×	×		×		A
10		×		×	×		A
11		×	×	×			A
12		×	×			×	A
13		×		×		×	A
Five independents							
14	×		×	×	×	×	R

TABLE 1. Overview of the models produced by the EH algorithm. The last column indicates the state of the model: Blank: intermediate model; U: kept by user intervention; A: accepted; R: rejected.

Swallowing (SWA) turned out to be a central variable: it occurs in all accepted models. The location of the tumor (MAL), although in itself related to the survival of the patient, seemed to be of less consequence. Minor complications (MIC) and Antigen-DR at followup (ANF) seem to be related: with hospitalization duration and major complications they form two groups of three variables that enter pairwise in the accepted models.

6 Discussion and Conclusion

The concept of methodological modelling allows us to arrive at a more precise description of the modelling process. However, it is necessary to make a distinction between modelling in a confirmative or an exploratory context, where 'context' refers to the nature of the substantive problem. In both cases it is clear that, although statistical modelling is a well-developed discipline, the two other parts of the methodological modelling process, (a) translation of the research problem into a specification of the space of statistical models and (b) the interpretation of the results have less received much less attention.

We use a diagramming method introduced in Adèr (1995) as a formalization method. It allows a well-defined graphical representations of methodological concepts. The diagrams can be automatically translated into a algebraic specification similar to the specification used to specify structural equation

models. This in turn allows an immediate formulation of an algorithmic representation.
As an example, the method of Edwards and Havránek to search a model space is used to demonstrate that this method allows higher conceptual interpretations eventually requiring user interaction. Apart from this, the method makes high collinearity between variables visible.
Table 1 and its interpretation suggest that in many modelling situations the 'optimal' model produced by conventional software is by no means the only model. Instead, an impression should be obtained of the *structure* of the model space and this kind of model search procedure can provide this structure.

A scrupulous analysis of the modelling process makes it clear that there is more to it than Statistics proper, although statistical modelling is an essential part of the process. Both the translation of subject matter notions into appropriate statistical models and the back-translation of the results of the statistical analysis in terms of the original research question deserve thorough analysis. In this article, we only made a beginning by precisely describe what is going on. The EH procedure was used to demonstrate that an analysis is possible that allows for a high-level conceptual interpretation.

Acknowledgement

The authors express their gratitude to Marjan van Bokhorst for making available the data set described in Section 3 to demonstrate the merits of the EH algorithm.

References

Adèr, H. J. (1995). *Methodological knowledge: Notation and Implementation in Expert Systems.* Phd thesis, University of Amsterdam.

Dobson, A. J. (1983). *Introduction to Statistical Modelling.* London New York: Chapman and Hall.

Edwards, D. (2000). *Introduction to Graphical Modelling* (second ed.). New York: Springer.

Edwards, D., & Havránek, T. (1987). A Fast Model Selection Procedure for Large Families of Models. *Journal of the American Statistical Association*, *82*(397), 205–213.

Van Bokhorst-de van der Schueren, et al.. (2000). Survival of malnourished Head and Neck Cancer Patients Can Be Predicted by Human Leukocyte Antigen-DR Expression and Interleukin-6/Tumor Necrosis Factor-à Response of the Monocyte. *Journ. of Parenteral and Enteral Nutrition*, *24*(6), 329–336.