Using hierarchical models to analyze facet-data

Joop Hox & Nijs Lagerweij
Department of Education, University of Amsterdam
Amsterdam, the Netherlands

Summary. If a facet design describes the question content as well as the respondent population, an analysis technique must be used that models the simultaneous and interaction effects of both question and person characteristics. A powerful method for such models is the hierarchical linear regression model. This chapter describes this model and gives an example of its use in the analysis of facet-data of a study of children's environmental behavior.

1. Introduction

Theory is defined by Guttman as an hypothesis of a correspondence between a definitional system for a universe of observations and an aspect of the empirical structure of these observations, together with a rationale for such an hypothesis. The definitional system, often given as a mapping sentence, may contain both content facets that define characteristics of the observation system (e.g. interview questions) and facets that define characteristics of the members of the population. If facet design is used to construct a survey questionnaire, the Cartesian product of the content facets defines a population of possible questions, and the Cartesian product of the person facets defines a stratified research population.

Facet design combines elements of experimental design with elements of survey sampling. The content facets are usually treated as fixed: the facets in the design contain all subcategories of interest. In practice, many facet studies employ a complex mapping sentence that generates too many questions than can be used in a single interview, and as a result only a

subsample of questions can be used. Respondents are most often sampled at random.

The person facets define variables that describe persons, and the content facets define variables that describe questions. Since in general each person will answer many questions, a facet design with both content and person facets produces hierarchical or multilevel data, with questions nested within persons.

The most popular analysis of facet data is to treat the persons as interchangeable. The responses on the common response range are used to compute a similarity matrix between the questions, and smallest space analysis is employed to produce a low-dimensional representation of these similarities. Finally, the geometric structure of this representation is interpreted in terms of the properties of the defining facets.

Hierarchical modeling of facet data takes a different viewpoint. The responses on the common response range are viewed as observations of what occurs when a specific person encounters a specific question. The goal of the analysis is to determine which question and person characteristics (as defined by the facet design) predict the outcome of this encounter. The analysis problem points to the application of a hierarchical regression model, with the response as criterion and the question and person facets as predictor variables.

2. The hierarchical linear regression model

For general references to the hierarchical linear regression model we refer to the monograph by Bryk and Raudenbush (1992); an application that resembles our analysis here is presented by Hox et al (1991). The hierarchical regression model for facet data consists of two parts: one modeling the effects of content facets within persons, and the other modeling

the effects of person facets between persons. The model includes interactions between content and person facets.

For each person, we have a within person regression equation of the form:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \ldots + \beta_{pj}X_{pij} + \varepsilon_{ij} \qquad (1)$$

where Y_{ij} is the response by person j to question i,
X_{pij} is the value of content facet variable p for person j and question i,
β_{pj} is the regression coefficient of content facet variable p for person j,
ε_{ij} is a residual error term.

Thus, each person has its own specific regression equation relating the content facet variables to the response variable. In the hierarchical linear regression model, these regression coefficients (intercept and slopes) are allowed to be different between persons. Thus, each of the regression coefficients in equation (1) is viewed as a stochastic variable that can in turn be modeled by a between-person regression equation of the form:

$$\beta_{pj} = \gamma_{p0} + \gamma_{p1}Z_{1j} + \ldots + \gamma_{pq}Z_{qj} + \delta_{qj} \qquad (2)$$

where β_{pj} is the within-person regression coefficient for content facet variable p for person j,
Z_{qj} is the value of person facet variable q for person j,
γp_q is the regression coefficient that models the effect of person facet variable q on the within-person regression coefficient β_{pj},
δ_{qj} is a residual error term.

Substituting (2) into (1) produces the single-equation expression of the model:

$$Y_{ij} = \gamma_{00} + \gamma_{p0}X_{pij} + \gamma_{0q}Z_{qj} + \gamma_{pq}Z_{qj}X_{pij} + \delta_{pj}X_{pij} + \delta_{oj} + \varepsilon_{ij} \qquad (3)$$

The hierarchical linear regression model is often presented as a hierarchical system of regressions as in equations 1 and 2. The single equation expression 3 shows that the hierarchical linear regression model is a regression equation involving both content facet and person facet variables and their interactions, with a complex error term. The regression part of the model that contains the regression coefficients gamma constitutes the fixed part of the model. The terms $\gamma_{p0}X_{pij}$ and $\gamma_{0q}Z_{qj}$ in the fixed part represent the direct effects of the content and person facets on the response, and the product term $\gamma_{pq}Z_{qj}X_{pij}$ that arises as a consequence of substituting (2) into (1) specifies interactions between content and person facets. The error term given by $\delta_{pj}X_{pij} + \delta_{oj} + \varepsilon_{ij}$ constitutes the random part of the model. The question level error term ε_{ij} has a normal distribution with expectation zero and variance σ_ε^2. The residual error terms $\delta_{pj}X_{pij}$ and δ_{oj} are assumed to be independent from the ε_{ij}, they have a multivariate normal distribution with expectation zero and (co)variances $\sigma_{pp'}^2$ collected in covariance matrix Ω.

An interesting property of the model is that makes it possible to partition the total between-person variance of the regression coefficients beta into systematic variance components given in Ω and residual sampling variance. The systematic variance in Ω represents the variance of the regression coefficients for the content facets that is not accounted for by the person facet variables in the model. In the 'intercept-only' model, which is the model without predictor variables, the ratio (systematic variance)/(total variance) for the intercept is the familiar intraclass correlation. In other models the variances are conditional upon the predictor variables in the model; the ratio (systematic variance)/(total variance) for a regression coefficient (intercept or slopes) is an estimate of the reliability of the estimates of the regression coefficients for the content facets (cf. Bryk and Raudenbush, 1992).

The available software for the hierarchical linear regression model uses an iterative maximum likelihood procedure for the parameter estimates. The estimation procedure also produces asymptotic standard errors for the

gammas and the variances, and a deviance that indicates the overall fit of the model. For the example presented below, we used the program HLM by Bryk and Raudenbush because it directly computes the reliability estimates mentioned above; with other programs (such as Longford's VARCL or Goldstein's ML3) the reliabilities must be hand-calculated. HLM also produces estimates of the content facet's regression coefficients for the individual persons, using a Bayesian procedure that weighs the individual person's regression coefficients by their reliability (Bryk and Raudenbush, 1992).

3. Example

The example data stem from a study of children's environmental behavior (for details see Lagerwey, forthcoming). Most social studies of environmental issues focus on environmental attitudes or awareness. The example study focusses on children's behavioral reactions in a wide variety of specific situations. The behavioral reactions concern possible choices between behavior that is detrimental to the environment in some respect (e.g. throwing an old battery in the garbage can) versus behavior that is more environment friendly (e.g. bringing old batteries to a chemical waste disposal center). Thus, the response range is a choice for behavior that is: *very environment-unfriendly ... very environment-friendly.* The item-domain was defined by a complex facet design. The theoretical background of the study is rational choice theory, which assumes that the behavior of individuals is guided by an evaluation of the costs and benefits of the expected consequences of a set of possible behaviors. This is incorporated in the facet design by including facets for the costs of the environment-friendly behavior. There are three facets that refer to the costs: one facet for *physical costs* (e.g. having to walk a large distance to bring the battery to a disposal center), and one facet for the *social costs* (going against the will of the social environment). The third cost-facet is nested within the social cost facet; if the social costs are high

they may be *explicit or implicit costs* (explicit costs meaning that there is a social referent physically present in the situation; e.g. environment-friendly behavior requires facing explicit ridicule from friends). The other facets were taken from the literature about consumer behavior and environmental studies. They are: *social environment* (e.g. behavior takes place at home), *initiative* (own or others'), and *consumption phase* (class of consumption activity). Thus, the item generating mapping sentence can be formulated as:

The choice by respondent (R) in a situation that is { Home / School / Other },

in { Buying / Using / Disposing } of a good by { Own / Others } initiative and

with { Low / Medium / High } physical costs and { None / Any { Implicit / Explicit } } social costs

—> environmentally { Unfriendly / . . . / Friendly } behavior.

Note that there is one unusual feature in the mapping sentence: the 'implicit/explicit social costs' are nested within the element 'high costs' of the social costs facet. The other facets define the questions by a complete

cartesian product. Thus, in this mapping sentence each question is defined by a structuple of up to six facets.[1]

Four respondent characteristics are also included in the design. Since the responses to our questions can be strongly influenced by social desirability bias, a *social desirability* scale for children is included in the questionnaire. For exploratory reasons, a second scale is included that probes the *social responsibility* in other domains than environmental behavior. Finally, the demographic variables *age* and *sex* are included as background variables. Ideally, these four characteristics should be included in the mapping sentence as facets describing the respondent (R). Since their inclusion at present is either for the purpose of control (social desirability) or purely exploratory (age, sex, social responsibility) we will not explicitly incorporate them in the mapping sentence.

The mapping sentence was used to generate a total of 90 questions. Since many questions closely resemble each other the children were presented with different subsets of 30 questions each. There are 15 different subsets of 30 items, following a complex rotating design.

The questions were assembled in a questionnaire that was given to a sample of 79 children between the ages of seven and thirteen. The 15 different versions of the questionnaire were distributed at random.

A hierarchical analysis looks upon the data as a sample of 2370 responses produced by a sample of 79 respondents. The six question and four person characteristics can be entered as predictor variables. From the general class of hierarchical regression models we can choose a number of models, either a priori based on theoretical considerations, or a posteriori based on statistical considerations such as significance. Since three of the six question facets were derived from an explicit theoretical base in rational choice theory, these predictors will always be included in the model. The other

[1]The study actually uses a more complex mapping sentence; our example is a subset from the study.

predictors will be tested for significance, and only be included in the final model if they meet the conventional .05 level. Table 1 below gives the results for a sequence of models:

Table 1. Results for environmental behavior data

Predictor variables	Intercept only	Fixed content slopes	Random content slopes	Random slopes + person vars.	Selected model
Fixed part					
content facets:					
intercept	3.92	2.90	2.85	-1.92	$-.37^{ns}$
soc.env.		.25	.26	.26	.26
cons. phase		.35	.36	.36	.36
initiative		.20	$.21^{ns}$	$.21^{ns}$	$.21^{ns}$
phys. cost		-.32	-.30	-.30	-.30
soc. cost		$.13^{ns}$	$.12^{ns}$	$.12^{ns}$	$.12^{ns}$
(in)direct soc. cost		-.56	-.56	-.56	-.56
person characteristics:					
age				$.09^{ns}$	
sex				.35*	.40
soc.des.				$.06^{ns}$	
soc.resp.				.12	.13
Random part					
σ_ε^2	4.34	4.08	3.72	3.72	3.72
σ_{00}^2(interc)	.84	.85	3.22	3.11	3.05
σ_{11}^2(soc.env.)			.13	.13	.13
σ_{22}^2(cons.ph.)			.35	.35	.35
σ_{33}^2(init.)			.46	.47	.47
deviance	9660.7	9556.6	3498.0	3494.0	9485.9

*marginal: p=0.08

The first model in Table 1 is the 'intercept only' model. This is useful as a baseline model. It decomposes the total population variance in (estimates of) the question level variance σ_ε^2 and the person level variance σ_{00}^2.

The second model introduces all content facets as fixed predictors. The social costs appear to be not significant, but because this is a theory-derived facet, we will not exclude this predictor from the predictor set.

The next model assumes that all regression slopes for the content facets are different for all persons. The slope variance for the three cost facets turned out to be insignificant. Thus, the third model in Table 1 specifies random regression slopes only for social environment, consumption phase, and initiative. In this model, the regression slope of initiative is no longer significant. However, we keep it in the model because this slope shows significant variation across respondents.

The last model in Table 1 includes the respondent variables. Only social responsibility has a significant effect.

4. Discussion

The general hierarchical regression model in equation 3 also includes cross-level interactions that may explain the slope variation between the respondents. In our example data, there were no significant cross-level interactions, and including such interactions in the model also did not reduce the between respondent slope variance. The interpretation of the parameter estimates in the fourth model is straightforward: children chose the environment-friendly behavior alternative more often when the behavior takes place not at home or school, in the disposal phase, and when costs are low. The social costs seem to count only when they are explicitly visible. In general, the responses are not influenced by social desirability. There is a relationship with socially responsible values in other domains, which suggests interesting relationships with conceptually related behavior domains where choices exist between behaviors with beneficial or detrimental consequences for society.

It is interesting to compare the results from the hierarchical regression model with the results that are typically produced by smallest space analysis (SSA). Each child has responded to a subset of 30 items, and can be said to have 60 missing values for the other items. We can compute a correlation

matrix using pairwise deletion and do a SSA. However, there are 4005 pairwise combinations of 90 questions, of which each respondent has received a subset of 439. Thus, each pairwise combination of questions is observed for an average of 9 respondents. Clearly, a smallest space analysis of this correlation matrix would analyze mostly sampling error.

Still, some of our results can be related to hypotheses more usually discussed in geometric terms. For instance, the second law hypothesizes that structuples that share a number of elements from different facets will be closer in the projection space than structuples that are dissimilar. In our regression equation, we predict for questions that have identical predictor profiles the same response. Thus, if we have an ordered facet that defines a direction in the projection space, we should find that effect as a significant regression slope in our hierarchical regression analysis. We do not mean that a hierarchical regression analysis is functionally equivalent tt a SSA. First, hierarchical regression analysis assumes linear relationships and multivariate normality; SSA assumes neither. Furthermore, circular structures are simply to depict in SSA and difficult to model in a regression model. A nominal facet that defines a circular structure can of course be modeled in a statistically correct manner by including dummy variables, in the regression analysis, but the analyst will loose sight of the fact that a simple geometric structure underlies the pattern of the slopes for the dummy variables. In our view, hierarchical regression analysis and SSA should be regarded as techniques that complement each other..

5. References

Bryk, A.S. & Raudenbush, S.W. 1992. **Hierarchical Linear Models**. Newbury Park: Sage.

Hox, J.J., Kreft Ita G.G. & Hermkens, P.L.J. 1991. The analysis of factorial surveys. **Sociological Methods & Research**, 19, 4, 493-510.

Lagerweij, N. (forthcoming). Environmental behavior of children.