Guttman-Rasch models in drug-use research

Jan de Leeuw
Department of Data Theory FSW
University of Leiden

The basic idea of these notes is to keep things simple, and still derive fairly complicated and possibly quite realistic models. Technical note: we use structural statistical models, not functional ones, because in drug research we are mainly interested in (sub)populations, and not in the precise involvement of individuals.

(a) Guttman model (deterministic)

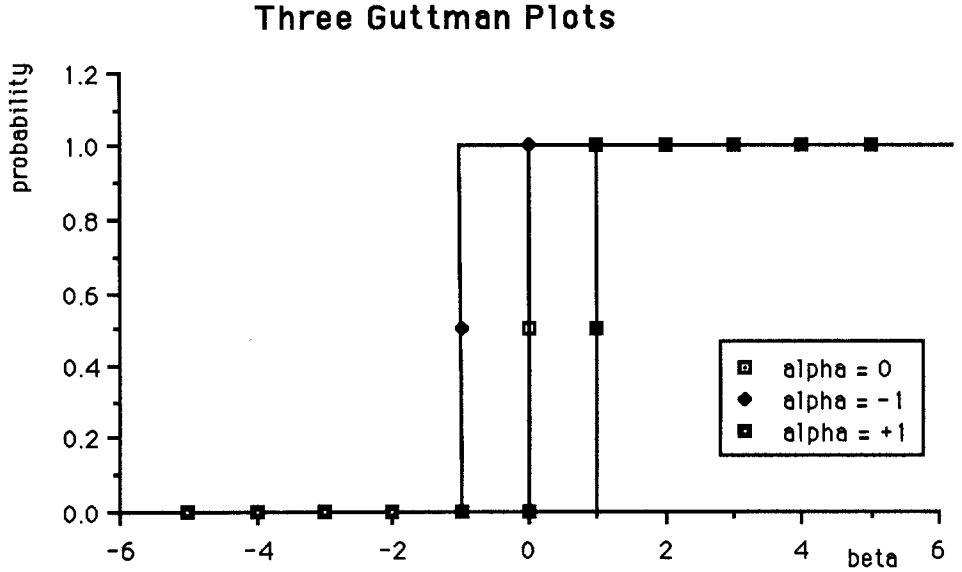
There are m drugs, and m corresponding parameter values α_j. Let us call α_j the seriousness of drug j. Individuals vary on the real line, we use β_i for the involvement of individual i. Suppose $X = \{x_{ij}\}$ is a data matrix with responses of the n individuals to the m items, here $x_{ij} = 1$ if i uses j and $x_{ij} = 0$ otherwise (or if i has used j, and so on). According to the Guttman model individual i uses drug j if $\beta_i > \alpha_j$ (i.e. if the individual dominates the drug), otherwise i does not use j (and the drug dominates the individual). This leads to a purely algebraic scaling problem. Given m drugs and n individuals, we want to find scales $\mathbb{a} = \{\alpha_i\}$ and $\mathbb{b} = \{\beta_j\}$ such that $x_{ij} = 1$ if and only if $\beta_i > \alpha_j$.

The deterministic Guttman model has three obvious disadvantages as a data analysis technique. It is never true in practice (except possibly in some very special cases with very few drugs) , it cannot be analyzed by the usual statistical methods because it is deterministic, and it only determines scales up to a monotone transformation. The remedy is to make the model less strict and, at the same time, probabilistic.

(b) Guttman model (probabilistic)

We now assume a probability structure which is very similar to the deterministic model above. Individuals are assumed to be a random sample from a population in which the distribution of involvement is given by the distribution function $F(\beta)$. Drugs are characterized by step-functions G_j, such that $G_j(\beta) = 0$ if $\beta \leq \alpha_j$ and $G_j(\beta) = 1$ if $\beta > \alpha_j$. Thus the characteristic G_j of drug j steps from zero to one at α_j. Compare Figure 1. The interpretation is that someone with involvement β has zero probability of using drug j if $\beta \leq \alpha_j$, and unit probability of using if $\beta > \alpha_j$. Secondly we assume that, given a certain level of involvement, the fact if individuals use or do not use various drugs is independent. Another way of saying this is that involvement is the only factor determining drug use. If we study a subpopulation of individuals all with the same involvement, the use of the

assumption. The probabilistic Guttman model is build out of the three assumptions unidimensionality, step-functions, and local independence.



Now some consequences. The probability that a randomly drawn member of the population uses drug j is equal to $F(\alpha_j)$. Because we can estimate probabilities of use, we can also estimate the $F(\alpha_j)$. This means that we can estimate $\mathbf{a} = \{\alpha_j\}$ is we know or assume what F is (for example standard normal, or rectangular). We see that the indeterminacy of the parameter estimates is still roughly the same as in the deterministic model: if F is unknown then only the order of the α_j can be recovered. Also we see that the probabilities of use of separate drugs do not make it possible to test the model. But probabilities of use of pairs of drugs can be used. If (j,l) is such a pair, then the Guttman model says that either $\alpha_j < \alpha_l$, in which case there are no individuals using l but not j, or the other way around. Thus if we make 2 x 2 tables for all pairs of drug, the Guttman model says that always at least one of the two off-diagonal cells of this table will be empty. Moreover which one of the cells is empty should correspond with a consistent ordering of the drugs. And similar predictions are implied for higher order frequencies. Each pattern of use in which someone uses a drug and but does not use one of the less serious drugs has probability zero.

Again it is clear that the model is very strong, and again it can not be tested by standard methods. In fact any data set in which a patterns occurs which is ruled out by the model will have likelihood zero, which means that standard tests will almost always reject (and never incorrectly so). We have gained very little, and we shall not take this model very seriously.

(c) Weak Guttman model.

We have seen that the Guttman model implies that one of the two off-diagonal cells in each of the twofold tables is empty. This implies that the correlations between the drugs have the form $r_{jl} = \tau_j/\tau_l$. Here $\tau_j = \sqrt{F(\alpha_j)} \Big/ \sqrt{(1 - F(\alpha_j))}$. This only describes the correlation coefficients above the

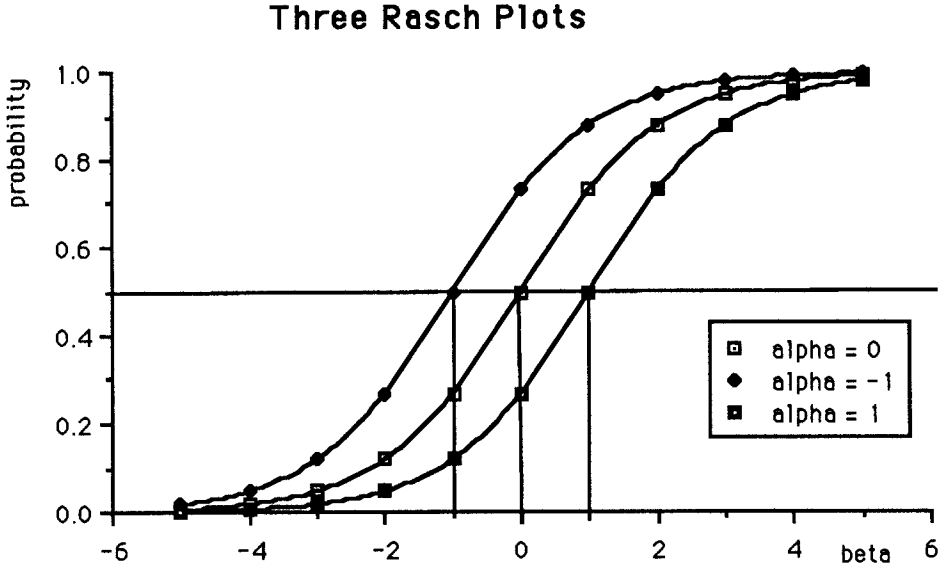
diagonal, for which $j < l$ and $\tau_j < \tau_l$. Thus the correlation matrix is a simplex. Any model which gives a correlational simplex is called a weak Guttman scale.

It is not difficult to test this necessary condition for a Guttman scale using (asymptotic distribution free) standard statistical theory for correlation structures. It is also not difficult to investigate if the scales $t = \{\tau_j\}$ of seriousness in different populations are indeed different. If we assume that the seriousness of the drugs is the same in all populations (which is in fact an identification condition in the weak Guttman model), then we test if the distribution of involvement differs. Observe that our hypothesis in framed in terms of a predefined order of seriousness.

(d) The Rasch model

We have seen that the Guttman-model was so strong because it combined the strong assumption of unidimensionality with the equally strong assumption of drug-characteristics which are step-functions. We now try to relax these assumptions, and see what we get.

In the Rasch model the assumption of unidimensionality is maintained, but a different and more smooth form of the characteristic is assumed. We suppose that seriousness and involvement are both real numbers, and we suppose that the probability that an individual with involvement β uses a drug with seriousness α_j is equal to $G_j(\beta) = 1 / [1 + \exp\{-(\beta - \alpha_j)\}]$. Compare Figure 2 for some Rasch characteristics. They increase smoothly from zero to one as a function of β. At the point $\beta = \alpha_j$ the characteristic is , which shows that the drug parameter can be interpreted as a treshold. Characteristics with different tresholds all have the same shape, but they are shifted along the involvement axis.



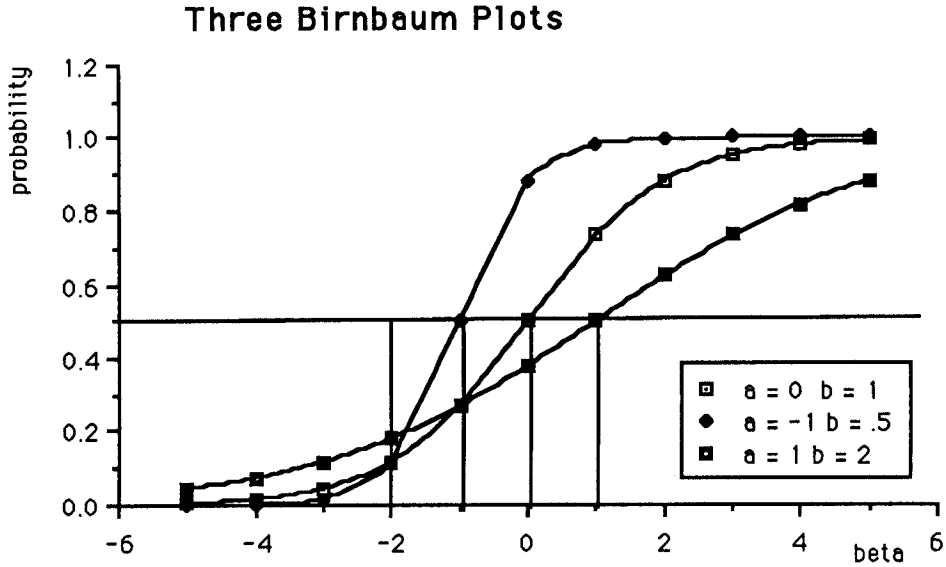
As in the Guttman models we assume a distribution of individuals $F(\beta)$, and we assume local independence. The statistical problem is to estimate the α_j and F, where we can *assume in addition*

that F has a certain parametric shape, for instance that it is normal. If we want to compare populations we can test if the tresholds are different, assuming that the distributions F are the same, or that the distributions are different assuming that the tresholds are the same. There are many intermediate possibilities, but the important thing is that usual multinomial large sample theory can be applied.

(e) Other unidimensional models

The Rasch model is more satisfactory than the Guttman model, because it does not assume the unrealistic step function for the characteristics, and because formal statistical treatment becomes possible. It is unsatisfactory, because the form of the characteristics is still very rigid, and because we still assume unidimensionality. This must be generalized in order to get more realistic models.

Let us first generalize the form of the characteristic. The first generalization is the two-parameter logistic model (also known as the Birnbaum model). It assumes that the characteristics are of the form $G_j(\beta) = 1 / [1 + \exp\{-(\beta - \alpha_j)/\sigma_j)\}]$. This means that there is not only a shift parameter α_j but also a scale parameter σ_j. The effect of scale is seen clearly in Figure 3. The slope of the characteristic (i.e. the discrimination of a drug) becomes more or less steep, and now two different characteristics intersect at a particular point. The intersection occurs at the value of β for which $(\beta - \alpha_j)/\sigma_j = (\beta - \alpha_1)/\sigma_1$, i.e. at $\beta = (\sigma_1\alpha_j - \sigma_j\alpha_1)/(\sigma_1 - \sigma_j)$. Intersection of characteristics means that the probability order of the drugs changes with involvement. For people who are very much involved heroin can be more probable than marihuana, for people who are just barely involved marihuana will be more probable. In the ordinary Rasch model this cannot occur: no matter how involved a person is, marihuana will always be the most probable drug of the two.



At this point it is convenient to introduce some special notation. We continue to use G_j for the characteristic of drug j, but we now use $\Psi(x)$ for the logistic function $1 / [1 + \exp(-x)]$. Thus for the Rasch model $G_j(\beta) = \Psi(\beta - \alpha_j)$ and for the Birnbaum model $G_j(\beta) = \Psi((\beta - \alpha_j)/\sigma_j)$. A very closely related model has characteristic $G_j(\beta) = \Phi((\beta - \alpha_j)/\sigma_j)$, where Φ is the cumulative standard normal curve. This can hardly be distinguished form the logistic curve. We call it the Lawley-model, because it was proposed by Lawley around 1940.

The Lawley model is equivalent to the following one-factor model. Suppose each drug j corresponds with a continuous normally distributed latent variable x_j, which can be interpreted as propensity to use drug j. Moreover suppose the x_j have a classical Spearman one-factor structure, with a normally distributed common factor ζ and with factor loadings σ_j. Suppose moreover that for each j there is a cutoff point α_j, where an individual uses a drug if his propensity to use it is larger than the cutoff point. This somewhat different formulation also turns out to be equivalent to the Lawley model. Observe that Lawley assumed not only a specific form for the characteristics, but also a normal distribution of involvement on the latent continuum. In our more general setup we can assume one without necessarily assuming the other.

Recent developments in psychometrics, which we shall not discuss in this introductory note, focus on developing nonparametric models. We suppose that $G_j(\beta) = G(\beta - \alpha_j)$, with G unknown, or even that all G_j are unknown. These developments are quite fascinating, but it is not clear yet if and in how far they really improve the more classical parametric techniques. In the rest of this note we shall concentrate on various multidimensional extensions of the models we have discussed so far.

(f) Population heterogeneity

A first trick which can be used, which in a sense still maintains the assumption of unidimensionality, is to use mixtures. We assume that the population we are sampling from is a mixture, with unknown mixing proportions, of two or more populations. Each of the subpopulations satisfies a unidimensional model, for example a unidimensional Rasch model. We estimate, simultaneously, the parameters of the Rasch models for the subpopulations and the mixing proportions.

This means that to describe a population consisting of p subpopulations (this number p can be compared with the dimensionality) we estimate p thresholds α_j, p involvement distributions F_j, and p mixing proportions π_j, which add up to one. Various restrictions come to mind: the thresholds can be the same and the distributions different, the distributions can be the same and the thresholds different, and so on. The distributions can again be parametric, i.e. restricted to be for instance normal, or they can be nonparametric. And the characteristics can be either Rasch, or Birnbaum, or Lawley, or also nonparametric. If we compare different populations we can test the hypothesis that they consist of the same subpopulations, but in different mixing proportions, and so on.

(g) Multidimensionality.

We get truly multidimensional model if we assume, from the start, that $G_j(\beta) = G[(\delta_j,\beta) - \alpha_j]$. Here both β and δ_j are p-dimensional vectors and (δ_j,β) is their inner product. The characteristic G can be either Φ or Ψ or nonparametric. Thus F must also be a p-dimensional distribution. We get a multidimensional Rasch model if $G = \Psi$ and all δ_j are the same, a multidimensional Lawley model if $G = \Phi$. The multidimensional Lawley model is equivalent, by the way, to the models for factor analysis of binary variables proposed by Muthen and Christofferson. Dimensionality can be tested in the usual way, and populations can also be compared in many ways.

(h) Conclusion

In drug-research, and in other situations in which we are interested in the one-dimensionality of a set of items, there are very little justification for continuing to use Guttman scales. They are inconvenient, both for rigourour computation and for statistical testing. Many probabilistic models are available which are far superior in these respects. The first candidates for testing in this context are the simple Rasch model, or the simple Lawley model. If they do not describe the data well enough, they can be generalized quite easily by using nonhomogeneous discrimination, nonhomogeneous populations, multidimensionality, and nonparametric drug-characteristics.

In this short note we did not consider variables with more than two states. Either one used a drug or one did not. All models discussed here have been generalized to multi-state ordered categorical data, but, not surprisingly, the model-choice problem and the computational problem become more serious for such variables. Moreover much larger samples are needed, because of the empty-cell problem. In cases in which there are not enough data we shall have to go to limited-information versions of these models, such as the weak Guttman model discussed above, which only use lower order marginals. In cases in which the data structure is even more complicated (mixed data, relatively few observations, non-random samples) the nonmetric data analysis techniques become far more useful.