Numeric scoring of ordered factors – a proposal

Adelchi Azzalini
University of Padua, Italy

11th Tartu Conference on Multivariate Statistics
26th June 2024

The question

- Consider a regression-type setting (linear, GLM, . . .)
- Some explanatory variables are ordered factors
- For example, 'quality' level can be expressed as
  very poor, poor, acceptable, good, very good
- Question: how to include them in the linear predictor?
- . . . a long-standing question!

Ordered factors

- Denote the levels of ordered factor F by $F_1, \ldots, F_K$
- Implicit assumption is that the levels have ordered effects:

$$f_1 < f_2 < \cdots < f_K$$

- WLG assume $<$ in the ordering.
- WLG take $K > 2$.

Classics recommendations – 1

Armitage (1955, Biometrics):

2. *A test based on scores*

To measure and test the significance of the trend in the p_i, a natural procedure is to allot a score x_i to the i-th column ($x_1 < x_2 < \cdots < x_k$), and to perform some sort of regression analysis of p on x. In addition

The calculations cannot be performed until the scores x_i have been chosen. In the absence of any *a priori* knowledge of the type of trend to be expected, it seems reasonable to choose the x_i to be equally-spaced, and it will often be convenient to have them centred around zero. This is the procedure advocated by Yates. Thus, for k columns, we should

Classics recommendations – 2

Graubard and Korn (1987, Biometrics):

The purpose of this paper is to demonstrate that the rank statistics can be poor choices for testing independence when the column margin is far from uniformly distributed; see also Mantel (1963, 1979). This is because of the well-known correspondence between the rank tests and tests using scores with midranks as the preassigned scores. Therefore, the notion that rank tests avoid the arbitrary choice of column scores is misleading. Our recommendations for testing independence in an ordered $2 \times K$ contingency table, which are similar to those given by Armitage (1955), are as follows:

(i) If possible, develop reasonable column scores based on the substantive meaning of the column categories, and use them in the analysis.

(ii) If no natural column scores are available, then consider using equally spaced column scores in the analysis.

(iii) Always examine the midranks as scores to make sure they are reasonable before using a rank test.

In practice

- Scores "based on the substantive meaning" hardly ever available.
- Equally spaced scores widely used in practical work, but...
- ...questionable because of arbitrary assumption on the effects
- Current method of choice:
  - use equally-spaced scores, and in addition...
  - include covariates with values of (orthogonal) polynomials
  - maximum polynomial degree is $K - 1$
  - retain as many polynomial terms as suited for the data
  - $\Rightarrow$ several numerical covariates are involved, in general

Example with 6 levels

Using R code

```
> contr.poly(6)
           .L         .Q         .C      ^4      ^5
[1,] -0.5976  0.5455 -0.3727  0.189 -0.063
[2,] -0.3586 -0.1091  0.5217 -0.567  0.315
[3,] -0.1195 -0.4364  0.2981  0.378 -0.630
[4,]  0.1195 -0.4364 -0.2981  0.378  0.630
[5,]  0.3586 -0.1091 -0.5217 -0.567 -0.315
[6,]  0.5976  0.5455  0.3727  0.189  0.063
```

Note: `.L` values are *equally-spaced, centred at* 0

Constructing numeric scores – a proposal

- Wanted: a *single* set of numbers $(x_1, \ldots, x_K)$ to represent the 'factor scores', avoiding the use of multiple covariates.
- E.g. map observations $(F_2, F_5, F_1, \ldots)$ to $X = (x_2, x_5, x_1, \ldots)$.
- Appeal: many practitioners like to work with numeric scores (basic scores $1, \ldots, K$ are often adopted, in spite of known drawbacks)
- Ideally from "substantantive meaning", yes . . . but rarely feasible.
- Goal: choose $(x_1, \ldots, x_K)$ best supported by the data
- End point: scores representing 'true' values of the factor level
- Points to notice
  - Condition of monotonicity: $x_1 < x_2 < \cdots < x_K$
  - Irrelevance of scale and location:

$$a + b\,x_1, \quad a + b\,x_2, \quad \ldots, \quad a + b\,x_K$$

constitute an equivalent vector.

Choosing the scores – a general practical scheme

- Consider a regression-type model, involving a response vector, a set of explanatory variables and associated parameters β.
- Focus of GLM's to ease exposition
- Model fitting obtained by minimizing the residual deviance $Q(\beta)$ with respect to β
- Design matrix is $(X, Z, W, \dots)$ where X is the column assigned to factor F (but not yet available)
- Building X: select a flexible parametric function $s(k, \theta)$ for mapping factor levels to real numbers

$$s(k, \theta) : F_k \to x_k, \qquad k \in \{1, \dots, K\}$$

- Now optimize $Q(\beta, \theta)$ with respect to (β, θ).
- For any given θ, minimization wrt β is efficient.

Choosing a mapping function – using quantiles

- Target: chooing the mapping

$$s(k, \theta) : F_k \to x_k, \qquad k \in \{1, \ldots, K\}$$

- Option A: using flexible parametric distributions
- Translate $\{1, \ldots, K\}$ to probabilities $\{\frac{1}{K+1}, \ldots, \frac{K}{K+1}\}$
- Choose $s(p, \theta)$ to be the quantile function of a distribution which depends on parameter θ
- Recall that location and scale of the distribution are irrelevant.
- Adopt a parametric family with high flexibility of the shape.
- Families of special interest: Tukey's g-and-h, Johnson's S_U and others alike.

Numerical illustration – Diamond pricing data, 1

Aim: tell diamond price from carat (numeric) and ordered factors (clarity: 8 levels, color: 7 levels, cut: 5 levels)

Set $\sqrt{\text{price}}$ to be the response

classical summary of the linear model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.194	0.75	1.60	0.11
carat	65.317	0.71	92.31	0.00
clarity.L	24.154	1.58	15.25	0.00
clarity.Q	-11.701	1.24	-9.42	0.00
clarity.C	3.610	1.24	2.90	0.00
color.L	-13.271	0.96	-13.76	0.00
color.Q	-1.907	0.89	-2.13	0.03
color.C	1.979	0.85	2.33	0.02
color^4	3.369	0.78	4.30	0.00
cut.L	1.882	0.85	2.21	0.03

Residual std. deviation: 6.74 on 527 degrees of freedom

Numerical illustration – Diamond pricing data, 2

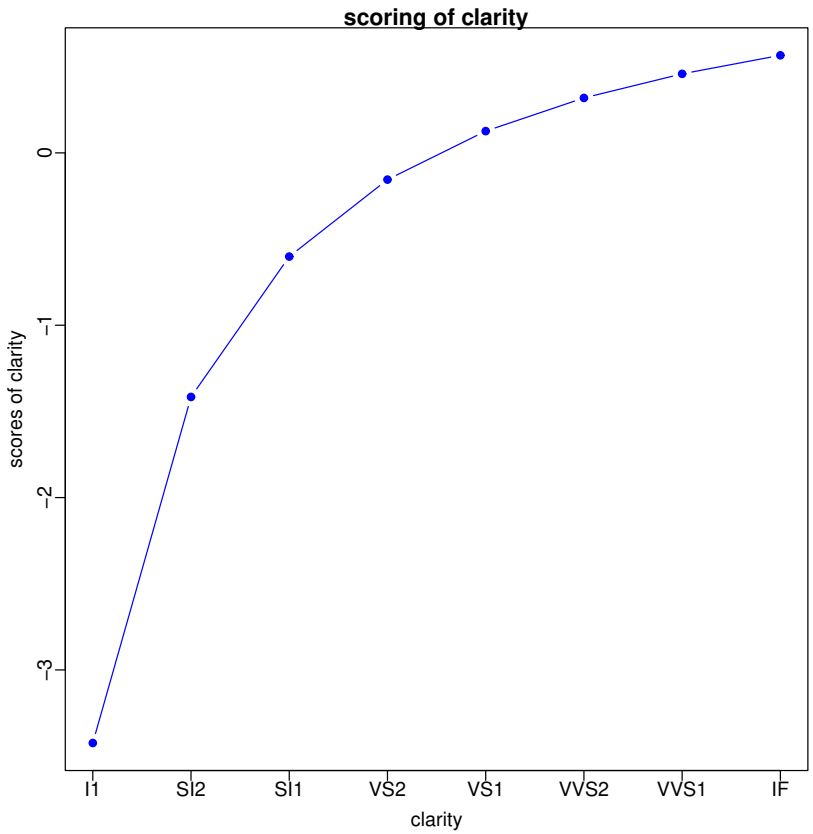
Summary of alternative model

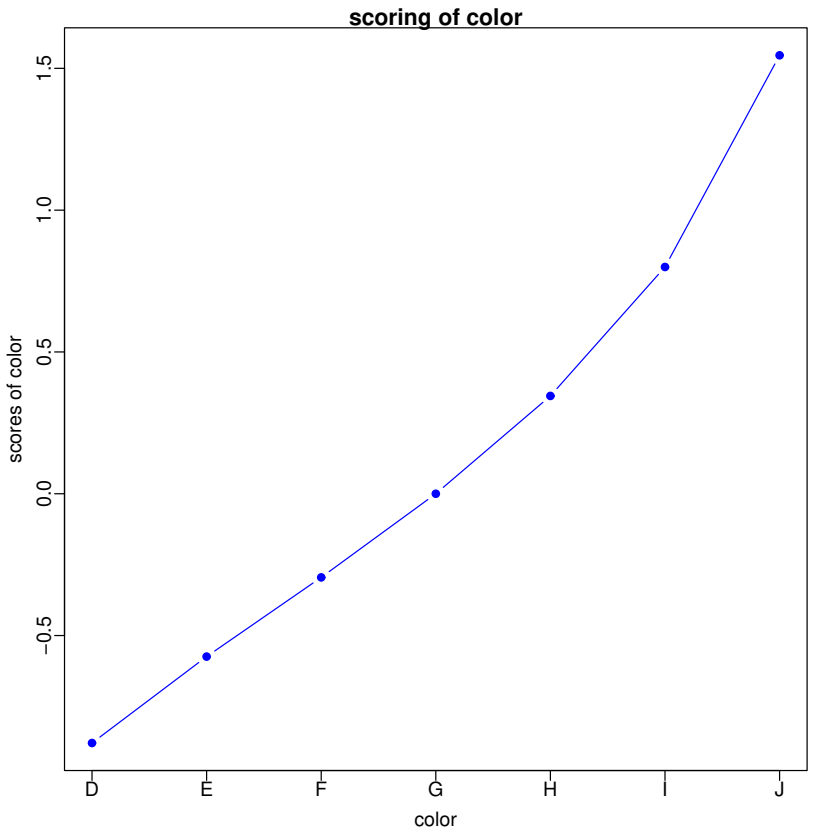
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.448	0.67	9.68	0.00
carat	65.001	0.71	91.03	0.00
clarity.score	8.618	0.47	18.52	0.00
color.score	-6.811	0.49	-13.98	0.00
cut.L	2.003	0.87	2.31	0.02

Residual std. deviation: 6.90 on 532 degrees of freedom

Numerical illustration – Diamond pricing data, 3

Numeric scores of 'clarity' and 'color' using g-and-h distributions (involves two parameters for each factor, hence four in total)





Some remarks

- Key feature is simplicity of interpretation
- Method based on explicit numeric scores of the levels
- Data fitting is analogous to the use of orthogonal polynomials

Choosing a mapping function – using splines

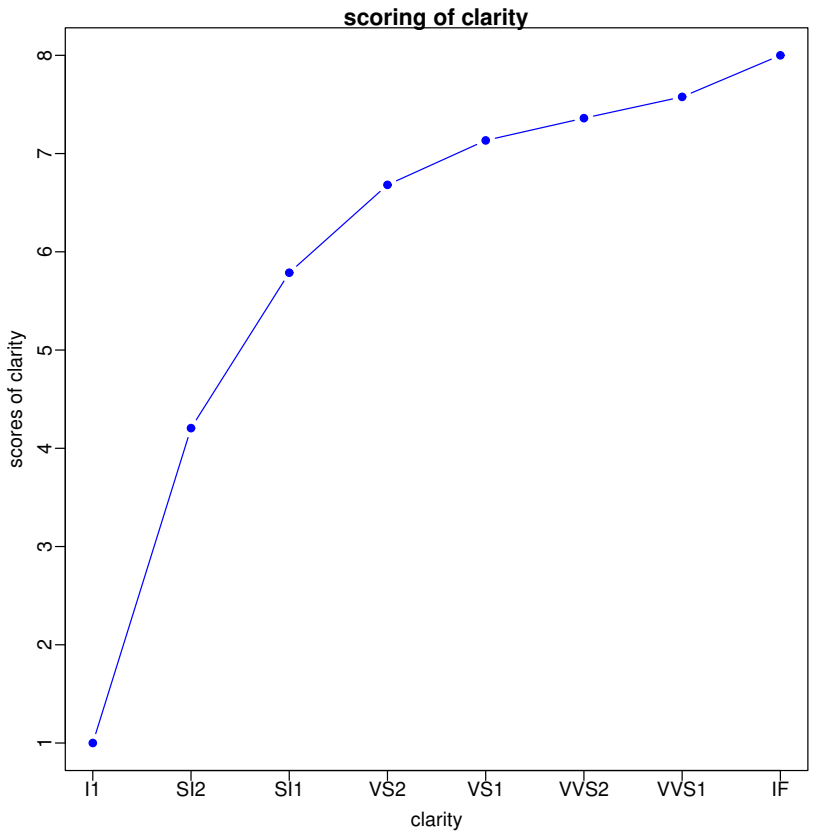
- Target: chooing the mapping

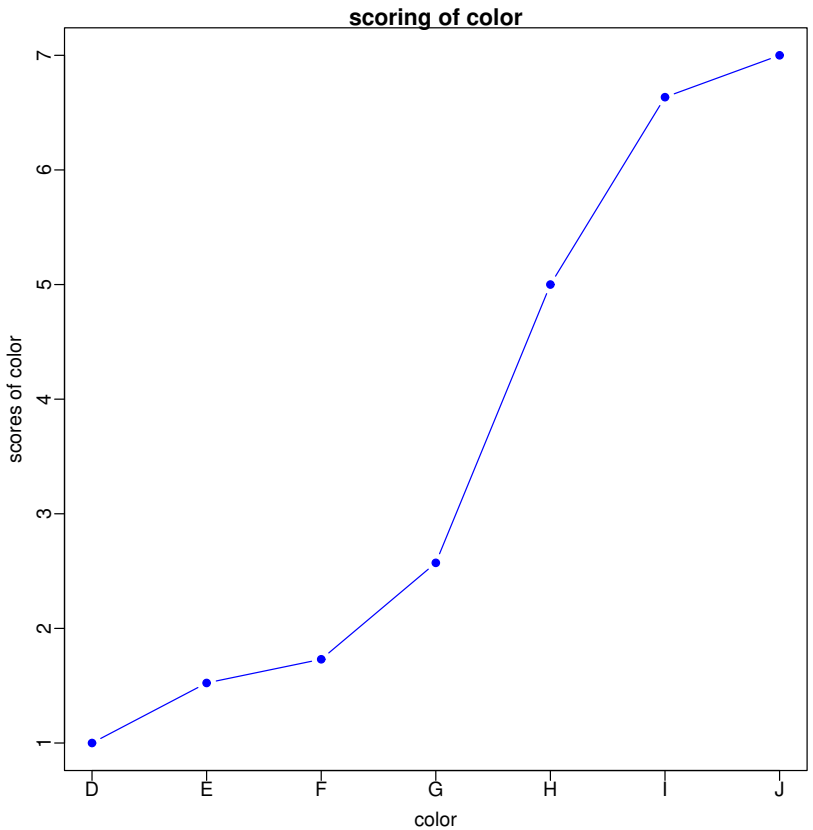
$$s(k, \theta) : F_k \to x_k, \qquad k \in \{1, \ldots, K\}$$

- Option B: use monotonic splines for $s(k, \theta)$
- Offers increased flexibility
- . . . even substantially, in case of several knots
- Price to pay: involves more parameters, two for each knot
- Technical detail: some of the knots must be suitably fixed

Numerical illustration – Diamond pricing data, 4

Numeric scores of 'clarity' and 'color' using splines
(using 1+2 knots, hence 6 parameters)





Numerical illustration – Diamond pricing data, 5

Summary of alternative model using splines

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-19.317	1.86	-10.40	0.00
carat	65.238	0.70	93.34	0.00
clarity.score	4.769	0.25	19.06	0.00
color.score	-2.242	0.15	-15.17	0.00
cut.L	1.870	0.85	2.20	0.03
Residual std.deviation: 6.74 on 532 degrees of freedom				

Recapitulate

- A method is proposed for assigning numeric scores to factors
- The main goal is to improve interpretability,
  while retaining good data fitting
- It allows to quantify the 'true' scores of the factor levels,
  a feature of interest for practical work

Resources

- Azzalini (2023). *Stat* 12, e624
  `DOI: 10.1002/sta4.624`
- follow-up paper at `https://arxiv.org/abs/2406.15933`
- R package `smof` on CRAN site