Some tools for the analysis of ordinal data

J.H. Randall
Department of Biometry, University of Stellenbosch, Stellenbosch, 7600 Republic of South Africa

Received 21 July 1992; accepted 10 December 1992

An error often committed in the analysis of data is described and the nature of the error is explained. A suitable method of analysis is identified and its use illustrated with two unusual examples.

'n Fout wat dikwels by data-ontleding begaan word, word uitgeken en die aard van die fout word verduidelik. 'n Geskikte metode van ontleding word genoem en die gebruik daarvan word met twee buitengewone voorbeelde geïllustreer.

Keywords: Generalized linear models, scales of measurement.

Introduction

A scientist sometimes makes measurements on an ordinal scale, for example a sheep may be 'very ill', 'ill', 'recovering' or 'completely recovered'. For convenience while performing the experiment, the scientist often records 1, 2, 3 and 4 in the place of these descriptions. Use of the abbreviations A, B, C and D would have been just as convenient but unfortunately the scientist is often misled by the use of numbers as symbols for the classes, with the result that the class labels are manipulated according to the rules of ordinary arithmetic. To understand why this is invalid (Stevens, 1946; 1958) one must appreciate that measurement classes have properties (i.e. they can be meaningfully manipulated in certain ways), as do numbers, but that one does not necessarily use all the properties of the number system when one uses a number as symbol for a measurement class. For example, the numbers 1, 2 and 3 are equally spaced; the 'distance' from 1 to 2 is that from 2 to 3. However, in the example, the distance from 'very ill' to 'ill' is not necessarily the same as that from 'ill' to 'recovering'. This explains why this sort of measurement is described as being on an ordinal scale. The particular property of numbers exploited on this scale is that $1 < 2 < 3$, corresponding with 'very ill' < 'ill' < 'recovering'. If the numbers had also reflected the distance between classes, the scale would have been interval.

Until recently, scientists wishing to analyse ordinal data were frustrated by the lack of suitable statistical methods. Consequently, they often proceeded to analyse their data by performing arithmetic directly on the numbers used to symbolize the measurement classes, i.e. by the methods appropriate to interval (or ratio) data. The invention of the concept of a 'generalized linear model' (GLM) by Nelder & Wedderburn (1972), and the subsequent flood of research in this area, should render this incorrect practice unnecessary. A very complete account of the theory concerned is to be found in McCullagh & Nelder (1989) while Dobson (1990) provides a quick introduction to the basic ideas of GLM techniques. The latter has little to say about ordinal data and should be read in conjunction with McCullagh (1980). An alternate school of thought is exemplified by Agresti (1984) and, perhaps most recently, by Lipsitz (1992).

The purpose of this paper is to draw attention to some techniques not described by any of the above publications.

Could an interval method of data analysis be used?

In the absence of an appropriate ordinal method of analysis, there is sometimes a need to know whether ordinal data can be treated as though it were interval data. The requirement to be met is whether the ordinal-scale classes can be treated as rounding intervals for an underlying continuous variable.

An excellent example is supplied by an ordinal scale for the thickness of hair defined by the Karakul Breeder's Association of Southern Africa (KBA) in 1982. Apparently to avoid the use of a delicate measuring instrument, a five-point scale is used in place of hair-thickness measurements in microns. A slightly modified version of the KBA scale (an ambiguity has been removed) is given in Figure 1.

A second representation of the scale is given in Figure 2. The symbol '[' shows the point on the continuous scale (known as the cut-off point) at which one measurement class is separated from another. The symbols 1 to 5 are convenient abbreviations for the measurement classes. Let z represent a number on the continuous scale, x a number on the ordinal scale.

The data available will be the frequencies with which the numbers 1 to 5 were recorded; write f_i for the frequency with which i was recorded, where $f_i > 0$ is required, and write $f_.$ for $f_1 + f_2 + \cdots$. Then, assuming that the distribution of the

Hair thickness	< 27	$27 — 30^-$	$30 — 35^-$	$35 — 38^-$	$\geqslant 38$
KBA scale	Thin	Medium thin	Medium	Medium thick	Thick

Figure 1 An example of how a continuous scale (hair thickness in microns) may be converted to an ordinal scale (KBA 5-point classification).

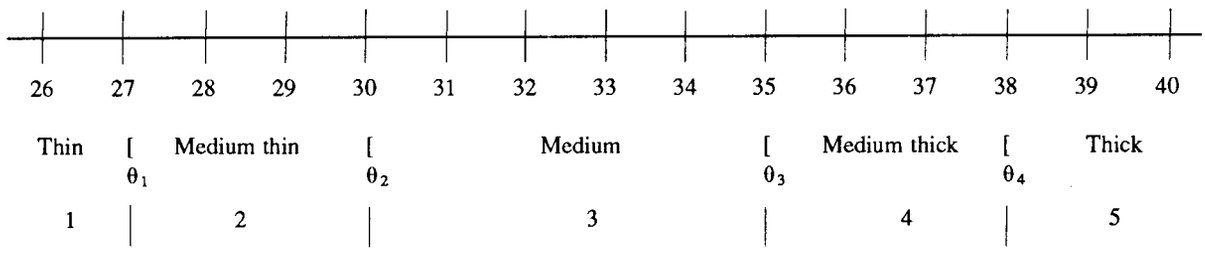

Figure 2 Defining an ordinal-scale variable on a continuous variable.

random variable Z is such that $P(Z < z) = F(z)$ for some function $F(\ )$ and writing θ_i for the cut-off values (naturally scaled, $\theta_1 = 27$, $\theta_2 = 30$, etc.), it follows that $f_1/f_.$ is an estimate of $F(\theta_1)$, $(f_1 + f_2)/f_.$ estimates $F(\theta_2)$ etc., and therefore that $F^{-1}(f_1/f_.)$ estimates θ_1, $F^{-1}([f_1 + f_2]/f_.)$ estimates θ_2, and so forth. The function $F(\ )$ is the *cumulative distribution function* (c.d.f.) of z and in GLM terminology, the inverse function $F^{-1}(\ )$ is the *link function* of a GLM. Amongst symmetric distributions, a popular choice for $F(\ )$ is the c.d.f. of the (standard) Normal distribution (so that the link function is the so-called probit link function) but the logistic link function has the advantage that while there is little difference between it and the probit, it is very much easier to evaluate; $F(z) = [1 + e^{-z}]^{-1}$, so that the inverse function is $z = log[F(z)/\{1 - F(z)\}]$ in which (throughout this paper) '*log*' refers to 'natural' logs.

A question one will want to answer with the available data is whether the ordinal-scale symbols are assigned in correspondence with the underlying scale. Writing $\theta_1 = \theta$, it follows that $\theta_2 = \theta + 3\delta$, $\theta_3 = \theta + 8\delta$ and $\theta_4 = \theta + 11\delta$, in which δ ensures proper scaling. Since one wishes to compare a model with four unknowns (the θ_i) to a model with two unknowns (θ and δ) it follows that one can fit a GLM yielding a *deviance* with 2 degrees of freedom (df). Deviances are distributed (approximately) as chi-square, giving one the necessary theoretical basis with which to choose between the two models. Writing $\mathbf{X}\beta$ for the model to be fitted,

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 3 \\ 1 & 8 \\ 1 & 11 \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \theta \\ \delta \end{bmatrix}$$

The fitting of the restricted model is complicated by the fact that $F^{-1}(f_1/f_.)$ and $F^{-1}([f_1 + f_2]/f_.)$ are correlated (f_1 being common) but the reader need not be concerned about such technicalities. Computer programs exist which take care of such details. The author can supply a program based on the SAS Institute's interactive matrix language procedure, PROC IML. This program can fit models other than those described in this paper. A very general personal-computer program is also available from the author.

Commercially-available programs capable of fitting a GLM to data are GLIM, Genstat and Minitab. See Hutchinson (1985) or Ekholm & Palmgren (1989) for information on the use of GLIM to fit models to ordinal data or Jansen (1988) in the case of Genstat.

If the simpler model fits as well as the more general model, then there is reason to believe that judges are correctly classifying hairs into the specified thickness classes. Under these circumstances, the available data may be analysed as though the measurement scale were interval. However, neither z nor x is suitable for this purpose. An appropriate variable, y, could be defined by the class medians, viz. $y = F^{-1}[\frac{1}{2}F(\theta)]$ for 'thin', $y = F^{-1}[\frac{1}{2}\{F(\theta) + F(\theta + 3\delta)\}]$ for 'medium thin', $y = F^{-1}[\frac{1}{2}\{F(\theta + 3\delta) + F(\theta + 8\delta)\}]$ for 'medium', and so forth.

The example above is unusual in the sense that an underlying continuous variable to the ordinal variable exists. More usually, such a variable is imaginary. In these circumstances, one can do little more than ask whether the variable x could be analysed as though it were an interval variable. The prerequisite for this is that the cut-off values be equally spaced. The procedure for testing this hypothesis is virtually identical to that of the example. The only difference is that the second column of $\mathbf{X}$ is replaced by 0, 1, 2, 3 · · · . It is better to use the variable y (the group medians) rather than x in cases where the end classes are open-ended.

It must be emphasized that the practice of analysing ordinal data as interval data is second best to analysis with an appropriate ordinal method, no matter how much simpler and easier the interval method is.

Are two judges consistent?

It is sometimes necessary to decide whether two (or more) judges (or the same judge on two or more occasions) are consistent. In this event, given f_{ij}, the frequency with which case i (judge, occasion, or whatever) assigned the j^{th} scale value, the cut-off points for case i are estimated by $F^{-1}(f_{i1}/f_.)$, $F^{-1}([f_{i1} + f_{i2}]/f_.)$, etc. On the other hand, if there is consistency, the best available common estimate of the cut-off values will be obtained by pooling the frequencies for the j^{th} scale value; writing $f_{.j} = f_{1j} + f_{2j} + \cdots + f_{rj}$ and $f_{..} = f_{.1} + f_{.2} + \cdots f_{.c}$, the estimates are $F^{-1}(f_{.1}/f_{..})$, $F^{-1}([f_{.1} + f_{.2}]/f_{..})$, $F^{-1}([f_{.1} + f_{.2} + f_{.3}]/f_{..})$, and so forth.

The comparison between these two models can again be made with the deviance. This test is unusual (in the analysis of ordinal data by GLM) in the sense that there is an explicit formula for the deviance. For the $r \times c$ table of f_{ij}'s with row totals $f_{i.}$, column totals $f_{.j}$ and overall total $f_{..}$ it is

$$D = 2\left[\sum_{i=1}^{r}\sum_{j=1}^{c} f_{ij}\log(f_{ij}/f_{i.}) - \sum_{i=1}^{r}\sum_{j=1}^{c} f_{ij}\log(f_{.j}/f_{..})\right]$$

which is approximately distributed as chi-square, with $(r - 1)(c - 1)$ df. Note that if the assumption $f_{ij} > 0$ is violated one can add 0.5 to every f_{ij}, but this author agrees with those who

counsel against this practice. The practice is invalid if either of $f_{i\cdot} > 0$ or $f_{\cdot j} > 0$ is violated. In the latter cases, the row or column concerned must be removed from the table. A column containing only zeros except for one non-zero entry should be pooled with an adjacent column and such pooling should be continued until all columns contain at least two non-zero entries. In very sparse data even this precaution may not be enough.

Other analyses

The methods above are given in some detail since they are not obvious examples of a GLM. In the case of a designed experiment with ordinal responses, the necessary theory may be found in the references quoted above, as well as in Jansen (1991) and Randall (1989).

More esoteric analyses are also possible. For example, a variety of discriminant and cluster analyses are described by Randall (1991).

Animal breeders needing to estimate (the equivalents of) variances and covariances for ordinal variables will probably find papers like those of Thompson (1990) and Schall (1991) a good starting point.

Acknowledgements

I thank the referees, the editor and Prof. I.M.R. van Aarde for their respective contributions in the preparation of this paper.

References

AGRESTI, A., 1984. Analysis of ordinal categorical data. Wiley, New York.

DOBSON, A.J., 1990. An introduction to generalized linear models. Chapman and Hall, London.

EKHOLM, A. & PALMGREN, J., 1989. Regression models for an ordinal response are best handled as nonlinear models. *GLIM Newsl.* 18, 31.

HUTCHINSON, D., 1985. Ordinal regression using the McCullagh (proportional odds) model. *GLIM Newsl.* 9, 9.

JANSEN, J., 1988. Using Genstat to fit regression models to ordinal data. *Genstat Newsl.* 21, 28.

JANSEN, J., 1991. Fitting regression models to ordinal data. *Biom. J.* 33, 807.

KARAKUL BREEDER'S ASSOCIATION OF SOUTHERN AFRICA, 1982. Beskrywing van Karakoellammers en die beoordeling van foto's. 24th Yearbook, p. 13.

LIPSITZ, S.R., 1992. Methods for estimating the parameters of a linear model for ordered categorical data. *Biometrics* 48, 271.

MCCULLAGH, P., 1980. Regression models for ordinal data (with discussion). *Jl R. statist. Soc. B.* 42, 109.

MCCULLAGH, P. & NELDER, J.A., 1989. Generalized linear models (2nd edn.). Chapman and Hall, London.

NELDER, J.A. & WEDDEDBURN, R.W.M., 1972. Generalized linear models. *Jl R. statist. Soc. A.* 135, 370.

RANDALL, J.H., 1989. The analysis of sensory data by generalized linear model. *Biom. J.* 31, 781.

RANDALL, J.H., 1991. Cluster analysis and discriminant analysis of ordinal data. Technical Report, Department of Biometry, University of Stellenbosch.

SAS INSTITUTE INC., 1989. SAS/IML® Software: Usage and reference, Version 6 (1st edn.). Cary, NC: SAS Institute Inc., 1989. 501 pp.

SCHALL, R., 1991. Estimation in generalized linear models with random effects. *Biometrika* 78, 287.

STEVENS, S.S., 1946. On the theory of scales of measurement. *Science* 103, 677.

STEVENS, S.S., 1958. Measurement and man. *Science* 127, 383.

THOMPSON, R., 1990. Generalized linear models and applications to animal breeding. In: Advances in statistical methods for genetic improvement of livestock Eds. Gianola, D. & Hammond, K., Springer-Verlag, Berlin, p. 312.