Statistica Neerlandica (2010) Vol. 64, nr. 1, pp. 125–128
doi:10.1111/j.1467-9574.2009.00432.x

On the number of categories in an ordered regression model

Philip Hans Franses*

Econometric Institute, Erasmus School of Economics, PO Box 1738, NL-3000 DR Rotterdam, The Netherlands

J. S. Cramer

Tinbergen Institute, University of Amsterdam, Roetersstraat, 101 SWB Amsterdam, The Netherlands

We show that there is no formal statistical testing method to support combining categories in a standard ordered regression model. We discuss practical implications of this result.

Keywords and Phrases: ordered regression model, number of categories.

1 Introduction and motivation

The ordered regression model (ORM) is frequently used in marketing, as is reflected by the fact that it is included in many commercial statistical packages. In this model the dependent variable is not continuous but takes J discrete and ranked values (see Mckelvey and Zavoina (1975) for an early reference and, for example, Franses and Paap (2001, Chapter 6) for a recent treatment. An example appears typically in questionnaires, when individuals are asked to indicate whether they Strongly Disagree, Disagree, are Indifferent, Agree or Strongly Agree with a certain statement. It is then the aim of the ORM to investigate which behavioral characteristics of the individuals can explain this classification.

Usually the number of discrete outcomes of the dependent variable is fixed from the outset. In questionnaires, J is often 5 or 7. In practice, it may happen that one or more of these outcomes may not be observed, like nobody answers 'Disagree'. In that case, one must construct an ORM for only those outcomes which occur. It may also happen that for one or more outcomes there are only a few observations. In that case, one may wonder whether an outcome category can be combined with another category. In a similar vein, one may have a continuously observed dependent variable like individual buying behavior in terms of dollar sales, but in the end, one might be interested only in understanding which variables explain low-volume,

*franses@few.eur.nl

Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

medium-volume and high-volume buyers. One may then want to construct an ORM instead of a standard regression model.

In the present paper, we show that an analyst can always reduce the number of outcome categories for practical considerations, but that there is no statistical test that can support this decision. Hence, this decision concerns a matter of convenience or taste.

2 Preliminaries

Consider the latent variable y_i^*, which measures the true but unobserved attitude or opinion of an individual i. Suppose for notational convenience that it depends on a single explanatory variable x_i, i.e.

$$y_i^* = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{1}$$

where ε_i usually obeys either the logistic or normal distribution. Furthermore, suppose that y_i^* is mapped onto an ordered categorical variable as

$$Y_i = 1, \quad \text{if } \alpha_0 < y_i^* \leq \alpha_1 \tag{2}$$

$$Y_i = j, \quad \text{if } \alpha_{j-1} < y_i^* \leq \alpha_j \quad \text{for } j = 2, \ldots, J-1 \tag{3}$$

$$Y_i = J, \quad \text{if } \alpha_{J-1} < y_i^* \leq \alpha_J, \tag{4}$$

where α_0 to α_J are unobserved thresholds. As the boundary values of the latent variable are unknown, one can set $\alpha_0 = -\infty$ and $\alpha_J = +\infty$. In sum, an individual i gets assigned to category j if

$$\alpha_{j-1} < y_i^* \leq \alpha_j, \quad j = 1, \ldots, J. \tag{5}$$

The ORM now becomes

$$\Pr[Y_i = j \mid X_i] = \Pr[\alpha_{j-1} < y_i^* \leq \alpha_j] \tag{6}$$

$$\Pr[Y_i = j \mid X_i] = \Pr[\alpha_{j-1} - (\beta_0 + \beta_1 x_i) < \varepsilon_i \leq \alpha_j - (\beta_0 + \beta_1 x_i)] \tag{7}$$

$$\Pr[Y_i = j \mid X_i] = F(\alpha_j - (\beta_0 + \beta_1 x_i)) - F(\alpha_{j-1} - (\beta_0 + \beta_1 x_i)), \tag{8}$$

for $j = 2, 3, \ldots, J-1$, and

$$\Pr[Y_i = 1 \mid X_i] = F(\alpha_1 - (\beta_0 + \beta_1 x_i)), \tag{9}$$

and

$$\Pr[Y_i = J \mid X_i] = 1 - F(\alpha_{J-1} - (\beta_0 + \beta_1 x_i)), \tag{10}$$

where F denotes the cumulative distribution function of ε_i. Obviously, α_1 to α_{J-1} and β_0 are not jointly identified. This is usually solved by imposing $\beta_0 = 0$, and hence the ORM reads as

$$\Pr[Y_i = j \mid X_i] = F(\alpha_j - \beta_1 x_i) - F(\alpha_{j-1} - \beta_1 x_i). \tag{11}$$

Clearly, the effect of the explanatory variable on y_i is not linear. For interpretation, one may therefore consider the odds ratio

$$\frac{\Pr[Y_i \leq j \mid X_i]}{\Pr[Y_i > j \mid X_i]} = \frac{F(\alpha_j - \beta_1 x_i)}{1 - F(\alpha_j - \beta_1 x_i)}. \tag{12}$$

For the Ordered Logit model, the natural logarithm of this odds ratio equals $\alpha_j - \beta_1 x_i$, see Franses and Paap (2001, p. 117). This result shows that the classification into the ordered categories depends only on the values of α_j. This essential difference with, for example, the log odds ratio for the multinomial logit model, already provides an insight that the results of Cramer and Ridder (1991) do not carry through for the ORM, as we will demonstrate in section 3.

3 Main result

Consider the two categories j_1 and j_2, where j_2 is above and adjacent to j_1, both containing several observations, and suppose that one contemplates to combine the observations into a single category j^*. The question is whether one can statistically test whether this combination is not rejected by the data.

The probability of having observations in the joint category j^* is equal to

$$\Pr[Y_i = (j_1, j_2) \mid X_i] = F(\alpha_{j_2} - \beta_1 x_i) - F(\alpha_{j_1 - 1} - \beta_1 x_i), \tag{13}$$

while the probabilities for the individual categories are

$$\Pr[Y_i = j_2 \mid X_i] = F(\alpha_{j_2} - \beta_1 x_i) - F(\alpha_{j_1} - \beta_1 x_i), \tag{14}$$

and

$$\Pr[Y_i = j_1 \mid X_i] = F(\alpha_{j_1} - \beta_1 x_i) - F(\alpha_{j_1 - 1} - \beta_1 x_i). \tag{15}$$

If there is no distinction between the two classes j_1 and j_2, then the assignment of observations is random, i.e.

$$\Pr[Y_i = j_1 \mid X_i] = \pi \Pr[Y_i = (j_1, j_2) \mid X_i] \quad \text{and} \quad \Pr[Y_i = j_2 \mid X_i]$$
$$= (1 - \pi) \Pr[Y_i = (j_1, j_2 \mid X_i].$$

In order to determine the likelihood of all N observations, one needs to estimate the parameter π. The maximum likelihood estimator of this parameter is, of course, the fraction of observations in category j_1 over the observations in the joint category j^*. However, under the null hypothesis, this estimator is equivalent to the estimator for the unknown threshold parameter α_{j_1}. In other words, under the null hypothesis, the observations have the same likelihood, whether the categories are combined or not. And hence a formal statistical test cannot be performed.

4 Implications

The absence of a formal statistical test for combining categories in an ORM means that where each outcome category gets observed, and one wants to reduce the model to consider, say, only $J - 1$ categories, this decision cannot be subjected to a statistical test. Naturally, this also holds for the case where one wants to assign the observations of one category to its two adjacent categories.

A second implication concerns a comparison of a standard regression model with an ORM. Suppose one has observed a continuous dependent variable y_i, which one aims to link with an explanatory variable. One may be interested in categories of this y_i variable, like low, medium and high, and suppose one wants to understand how this categorization can be explained by the variables. One way to proceed now is to define these categories and use an ORM right away. A question could then be whether the standard linear regression would be better than the ORM or the other way around. The results in this paper suggest that a formal test is not possible.

We should remark that adding categories has an effect on efficiency. When there are more categories, one has more information about the regression line, and hence efficiency increases. Brant (1990) uses this notion for his Hausman-type test.

Acknowledgements

We thank Richard Paap for helpful comments.

References

Brant, R. (1990), Assessing proportionality in the proportional odds model for ordinal logistic regression, *Biometrika* **46**, 1171–1178.

Cramer, J. S. and G. Ridder (1991), Pooling states in the multinomial logit model, *Journal of Econometrics* **47**, 267–272.

Franses, P. H. and R. Paap (2001), *Quantitative models in marketing research*, Cambridge University Press, Cambridge.

Mckelvey, R. D. and W. Zavoina (1975), A statistical model for the analysis of ordinal level dependent variables, *Journal of Mathematical Sociology* **4**, 103–120.

Received: June 2009, Revised: July 2009.