On the number of categories in an ordered regression model

Philip Hans Franses*
Econometric Institute
Erasmus University Rotterdam

Mars Cramer
Tinbergen Institute
University of Amsterdam

March 28, 2002

Econometric Institute Report 2002-15

Abstract

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

Key words: Ordered regression model, number of categories

*We thank Richard Paap for helpful comments. Address for correspondence: Philip Hans Franses, Econometric Institute H11-34, Erasmus University Rotterdam, P.O.Box 1738, NL-3000 DR Rotterdam, The Netherlands, e-mail: franses@few.eur.nl

1 Introduction and motivation

The ordered regression model [ORM] is frequently used in economics, finance, marketing, psychology and sociology, 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, however, one or more of these outcomes may not be observed. 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, medium-volume and high-volume buyers. One may now wonder whether it would perhaps be better to construct an ORM instead of a standard regression model, for example to mitigate the effects of outliers.

In the present paper we show that a researcher can always reduce the number of outcome categories for practical considerations, but that there is no statistical test that might support this decision. In other words, the results in Cramer and Ridder (1991) for the multinomial logit model do not carry over to the ORM.

2 The model

Consider the latent variable y_i^*, which measures the genuine but unobserved attitude or opinion or an individual i. Suppose for notational convenience that it depends on

a single explanatory variable x_i, that is,

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

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

$$\begin{aligned} Y_i &= 1 \quad \text{if} \quad \alpha_0 < y_i^* \leq \alpha_1 \\ Y_i &= j \quad \text{if} \quad \alpha_{j-1} < y_i^* \leq \alpha_j \quad \text{for } j = 2, \ldots, J-1 \\ Y_i &= J \quad \text{if} \quad \alpha_{J-1} < y_i^* \leq \alpha_J, \end{aligned} \tag{2}$$

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

$$\alpha_{j-1} < y_i^* \leq \alpha_j, \ j = 1, \ldots, J. \tag{3}$$

The ORM now becomes

$$\begin{aligned} \Pr[Y_i = j|X_i] &= \Pr[\alpha_{j-1} < y_i^* \leq \alpha_j] \\ &= \Pr[\alpha_{j-1} - (\beta_0 + \beta_1 x_i) < \varepsilon_i \leq \alpha_j - (\beta_0 + \beta_1 x_i)] \\ &= F(\alpha_j - (\beta_0 + \beta_1 x_i)) - F(\alpha_{j-1} - (\beta_0 + \beta_1 x_i)), \end{aligned} \tag{4}$$

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

$$\Pr[Y_i = 1|X_i] = F(\alpha_1 - (\beta_0 + \beta_1 x_i)), \tag{5}$$

and

$$\Pr[Y_i = J|X_i] = 1 - F(\alpha_{J-1} - (\beta_0 + \beta_1 x_i)), \tag{6}$$

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

$$\Pr[Y_i = j|X_i] = F(\alpha_j - \beta_1 x_i) - F(\alpha_{j-1} - \beta_1 x_i). \tag{7}$$

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

$$\frac{\Pr[Y_i \leq j|X_i]}{\Pr[Y_i > j|X_i]} = \frac{F(\alpha_j - \beta_1 x_i)}{1 - F(\alpha_j - \beta_1 x_i)}. \tag{8}$$

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 the next section.

3 Reducing categories

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

$$\Pr[Y_i = (j_1, j_2)|X_i] = F(\alpha_{j_2} - \beta_1 x_i) - F(\alpha_{j_1 - 1} - \beta_1 x_i), \tag{9}$$

while the probabilities for the individual categories are

$$\Pr[Y_i = j_2|X_i] = F(\alpha_{j_2} - \beta_1 x_i) - F(\alpha_{j_1} - \beta_1 x_i), \tag{10}$$

and

$$\Pr[Y_i = j_1|X_i] = F(\alpha_{j_1} - \beta_1 x_i) - F(\alpha_{j_1 - 1} - \beta_1 x_i). \tag{11}$$

If there is no distinction between the two classes j_1 and j_2, then the assignment of observations is random, that is, $\Pr[Y_i = j_1|X_i] = \pi \Pr[Y_i = (j_1, j_2)|X_i]$ and $\Pr[Y_i = j_2|X_i] = (1 - \pi) \Pr[Y_i = (j_1, j_2|X_i]$.

In order to determine the likelihood of all N observations, one needs to estimate the above π parameter. The ML 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 Practical 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, then 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. However, in the end one is interested in categories of this y_i variables, like low, medium and high, and one wants to understand how this categorization can be explained by the variables. One way to proceed is now 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.

References

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: Cambridge University Press.

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.