Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Sciences 176 (2006) 465–474
INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL
www.elsevier.com/locate/ins

# Regression for ordinal variables without underlying continuous variables

Vicenç Torra [a,*], Josep Domingo-Ferrer [b], Josep M. Mateo-Sanz [c], Michael Ng [d]

[a] *Institut d'Investigació en Intel·ligència Artificial-CSIC, Campus UAB s/n, E-08193 Bellaterra, Catalonia, Spain*
[b] *Universitat Rovira i Virgili, Dept. of Comp. Eng. and Maths, Av. Països Catalans 26, E-43007 Tarragona, Catalonia, Spain*
[c] *Universitat Rovira i Virgili, Statistics and OR Group, Av. Països Catalans 26, E-43007 Tarragona, Catalonia, Spain*
[d] *Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong*

**Abstract**

Several techniques exist nowadays for continuous (i.e. numerical) data analysis and modeling. However, although part of the information gathered by companies, statistical offices and other institutions is numerical, a large part of it is represented using categorical variables in ordinal or nominal scales. Techniques for model building on categorical data are required to take advantage of such a wealth of information. In this paper, current approaches to regression for ordinal data are reviewed and a new proposal is described which has the advantage of not assuming any latent continuous variable underlying the dependent ordinal variable. Estimation in the new approach can be implemented using genetic algorithms. An artificial example is presented to illustrate the feasibility of the proposal.


* Corresponding author. Tel.: +34 93580 9570; fax: +34 93580 9661.
*E-mail addresses:* vtorra@iiia.csic.es (V. Torra), josep.domingo@urv.net (J. Domingo-Ferrer), josepmaria.mateo@urv.net (J.M. Mateo-Sanz), mng@math.hkbu.edu.hk (M. Ng).

0020-0255/$ - see front matter 
doi:10.1016/j.ins.2005.07.007

*Keywords:* Categorical variables; Ordinal scales; Linear models; Regression models

## 1. Introduction

Thanks to the existing computational power, the ease of storage and the availability of computer-based tools, companies and institutions gather and store huge quantities of data from customers, suppliers, users or internal processes. To exploit these huge sets (e.g., to extract relevant knowledge from them), analytical software tools are required. Data analysis and data mining are fields devoted to the construction and study of such tools.

In general, information is either represented by means of numerical or categorical data. In the numerical case, variables take values in a continuous domain. In the categorical case, variables take values in nominal scales (no comparison between categories is possible), ordinal scales (scales where categories are totally ordered), partially ordered scales (with a partial order in the domain of categories), etc.

While model building for continuous data is a stable and classical matter (numerical regression models are broadly known and used), analyzing and building models for categorical data is far less unified. Indeed, existing techniques for continuous data analysis are not easily exported to the categorical case. This is so because not all operations that can be carried out on numbers have their counterpart on categories: this is the case for arithmetical operations used in computing regression for continuous variables.

### 1.1. Contribution and plan of this paper

This work deals with model building for ordinal data. Existing approaches [1] either do not take ordinality into account or they assume that there is a latent continuous variable underlying the dependent ordinal variable. We present an extension of least-squares regression to ordinal data which, rather than assuming it, *builds* an optimal numerical mapping between the categories of the dependent variable and those of the independent variables.

The approach presented in this paper is especially appropriate for data mining and model building based on survey data. Indeed, a substantial part of the data collected from citizens by national statistical offices are ordinal but do not have an obvious numerical interpretation (see [3]).

Section 2 reviews existing approaches to regression on categorical data and highlights their weak points when applied to ordinal data. Section 3 introduces the notation used in the rest of the paper and gives some background on linear and non-linear regression for numerical variables. Section 4 presents our approach to regression model building for ordinal data. Section 5 describes

an application of our technique to a toy example with 3 variables and 14 objects. Section 6 contains some conclusions and suggestions for future work.

## 2. State of the art on regression models for ordinal data

When dependent variables are measured on an ordinal scale, there are many options to build a model. These include:

- Ignoring the categories of the variable and treating it as nominal, i.e. using Mlogit techniques (see [8] for an overview). The key problem here is loss of efficiency. Ignoring the fact that the categories are ordered means not using some of the information available, which may lead to estimating more parameters than necessary. Even if parameter estimates should still be unbiased, there is a high risk of obtaining non-significant results.
- Treating the dependent variable as though it were continuous (numerical). In this case, categories in the ordinal scale are numbered consecutively and plain least-squares regression is used as described in Section 3.1. This easy and common choice is reasonable when the dependent variable has a large number of categories (5 or more). In case of doubt, some truly ordinal approach (see below) should be used to confirm that the continuity assumption for the dependent variable does not result in significant distortions.
- Treating the variable as though it were measured on an ordinal scale, but the ordinal scale represented crude measurements of an underlying continuous variable. For example, the categories "Cold, Cool, Warm, Hot" can be seen as rough measures of a continuous variable Temperature. In this case, an ordered logit model such as PLUM can be used [7].
- Considering the dependent variable as measured on a true ordinal scale. This may be the most sensible option for ordinal variables which cannot be regarded as crude measurements of an underlying continuous variable. This is the case for professional rank, e.g. "Parson, Bishop, Archbishop". There is a lack of regression models that can take advantage of ordinality without assuming underlying continuous variables. The purpose of this paper is precisely to propose one such model.

## 3. Notation and background

In this section, we describe the notation used throughout this paper, which follows the one in [2]. Also, least-squares regression for numerical data is briefly recalled for later use.

We assume a two-dimensional table where one dimension corresponds to the set of objects (denoted by $\mathbf{O} = \{o_1, \ldots, o_M\}$) and the other dimension corresponds to the set of variables (denoted by $\mathbf{V} = \{V_0, V_1, V_2, \ldots, V_N\}$). For each pair (*object, variable*), the table contains the value that *object* takes for *variable*. The table can be modeled as a function

$$\mathbf{V} : \mathbf{O} \rightarrow D(V_0) \times D(V_1) \times D(V_2) \times \cdots \times D(V_N)$$

where $D(V_i)$ corresponds to the range of $V_i$ (we denote by $|D(V_i)|$ the cardinality of $D(V_i)$).

For simplicity, and without loss of generality, the $N$-dimensional function $V$ can be split into $N$ one-dimensional functions ($V_i(\cdot) : \mathbf{O} \rightarrow D(V_i)$) that assign a value for variable $V_i$ to a given object. Equivalently, $\mathbf{V}$ is assumed to be of the form

$$\mathbf{V}(O) = (V_0(O), V_1(O), V_2(O), \ldots, V_N(O)) \tag{1}$$

with the representation above, building a model means defining a mapping between the ranges of variables. Without loss of generality, we assume in what follows that the dependent variable is $V_0$. In this case, building a model is finding a function

$$\mathscr{F} : D(V_1) \times D(V_2) \times \cdots \times D(V_N) \rightarrow D(V_0)$$

such that $\mathscr{F}(V_1(O), \ldots, V_N(O))$ is similar to $V_0(O)$. We will denote by $\hat{V}_0(O)$ the estimation of $V_0(O)$ (i.e., $\hat{V}_0(O) = \mathscr{F}(V_1(O), \ldots, V_N(O))$).

### 3.1. Linear regression for continuous data

When $\mathscr{F}$ is restricted to be a linear model, we have that

$$\hat{V}_0(O) = \beta_1 V_1(O) + \ldots + \beta_N V_N(O)$$

In this case, building the model requires determining parameters $\beta_i$ for $i = 1, \ldots, N$. A way to do this is to use the least squares method. This is "to find the model output with the minimal sum of squared error loss function value" [10]. Using matrices and vectors, the solution of that minimization problem can be expressed by:

$$\beta = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V_0} \tag{2}$$

where $\beta = \{\beta_1, \ldots, \beta_N\}$, $\mathbf{V}_0 = \{V_0(o_1), \ldots, V_0(o_M)\}$ and where $\mathbf{X}$ is the matrix $\mathbf{V}$ without the column corresponding to variable $V_0$, i.e. $X = \{V_1, \ldots, V_N\}$.

### 3.2. Non-linear regression for continuous data

The model (2) is linear with respect to the variables considered. Even if the relationship between variables is known to be non-linear, it can be linearized by

using transformations on variables. A typical example is using log or exp transformations to turn an exponential or a logarithmic regression into a linear one. As [10] points out, "intelligent preprocessing can often reduce the complexity by simplifying non-linear to linear optimization problems".

## 4. A new approach for ordinal regression

Let us now consider the process of building a model for categorical variables in ordinal scales, i.e. a model for domains where elements can be compared but not operated. We want a model taking ordinality into account but not assuming any continuous variable underlying the ordinal variables. Our proposal builds on the least-squares regression for continuous data recalled in Section 3.1.

Two main difficulties appear when trying to use a continuous model on ordinal data:

(1) Addition and multiplication are not defined on ordinal scales.
(2) The mapping of the ordinal scale into a continuous domain is not unique.

The first difficulty is that least-squares regression as described above cannot be used directly on ordinal data. Arithmetical operations are not possible on ordinal data and alternative operators are needed. However, the only well-known and accepted operators for ordinal scales are the minimum and maximum. To avoid the definition of new categorical operators (as in e.g., [5]), we propose to map ordinal data into the unit interval $[0,1]$ and then use least squares regression on that interval. Let us denote the mapping of variable $V_i$ into $[0,1]$ by $f_i$.

The second difficulty is how to determine the mapping $f_i$ for each variable $V_i$. This is not straightforward. For each variable $V_i$, there exist several mappings $f_i: D(V_i) \to [0,1]$. If the domain of variable $V_i$ consists of $k$ categories $D(V_i) = \{c_1, \ldots, c_k\}$, the simplest definition is $f_i(c_j) = j/|D(V_i)|$. This definition (as any other possible definition) adds, as it were, shape to the domain, i.e. it fixes the difference between two consecutive categories. Consequently, the particular mapping chosen shapes (and strongly biases) the final regression model.

Since we are not assuming any continuous variable underlying the ordinal ones, it would be especially awkward to let the numerical mapping $f_i$ bias the resulting model. Our approach is not to fix the mapping beforehand but to include the estimation of the best mapping functions $f_i$ in the model construction. According to this, the problem to be solved can be formulated as follows:

**Definition 1.** Let **V** be the data as defined in Section 3. Building the model can be defined as finding $\beta_i$ and $f_i$ that minimize the following expression:

$$\sum_{j=1,M}\left(\sum_{i=1,N}\beta_i f_i(V_i(o_j)) - f_0(V_0(o_j))\right)^2 \tag{3}$$

It is important to note that, once the $f_i$ are known, coefficients $\beta_i$ are determined as described in Section 3.1.

### 4.1. Optimization using genetic algorithms

To solve the optimization problem stated in Definition 1, we use genetic algorithms. Our approach follows the type of genetic algorithm given in Appendix B of [6]. For a more detailed and comprehensive description of genetic algorithms, see [9,4].

We use the fact that the selection of the mappings $f_i$ induces the vector $\beta$. According to this, the problem can be reduced to finding functions $f_i$ that minimize Expression (3) (with $\beta$ computed as in Expression (2)).

Genetic algorithms require possible solutions to be encodable as chromosomes, i.e. binary strings. According to the above discussion, the only objects to be represented in a chromosome are the functions $f_i$. In the following subsections, we describe how these functions are encoded into chromosomes and what is the fitness function applied for evaluating chromosomes and guiding towards the optimal solution.

#### 4.1.1. Coding

In our case, each set of functions $\{f_i\}_{i\in\{1,\ldots,N\}}$ is a possible solution. We encode each function $f_i$ as a vector of $|D(V_i)|+1$ integers $(n_1,\ldots,n_{|D(V_i)|+1})$. From such a vector, $f(c_i)$ is defined as:

$$f(c_i) = \frac{\sum_{j\leqslant i} n_j}{\sum_{j=1,|D(V_i)|+1} n_j}$$

Note that this expression defines a monotonic mapping into $[0,1]$, but it is not necessary that, for the largest category $c_{|D(V_i)|}$, the equality $f_i(c_{|D(V_i)|}) = 1$ holds. In general, we are not interested in $f_i(c_{|D(V_i)|}) = 1$ because this would force all variables to have the same numerical domain. We allow different variables to map their categorical domain into different numerical domains (these being subsets of $[0,1]$).

Vector $(n_1,\ldots,n_{|D(V_i)|+1})$ is translated into a binary representation to obtain a chromosome (each integer component is translated into its corresponding binary equivalent).

#### 4.1.2. Fitness function

Application of genetic algorithms requires a function (the fitness function) to evaluate each chromosome. This function is defined using Expression (3).

However, as this latter expression is to be minimized and the fitness function is usually to be maximized, we define the fitness function as

$$1/\text{Expression} \quad (3)$$

In the above expression, $\beta_i$ corresponds to the parameters of the linear model computed using Expression (2).

*4.1.3. Genetic operators*

Genetic algorithms are an iterative technique, where in each iteration the best chromosomes of the population (the set of functions with a better linear model) survive and are the basis of the next population. Here, "best chromosomes" mean chromosomes that evaluate best in relation to the fitness function.

In our case, the next population is built by random selection of chromosomes from the previous population. The probability of selecting a chromosome $\text{chr}_i$ is proportional to the fitness of this chromosome. Equivalently, the probability of selecting $\text{chr}_i$ is $\text{fitness}(\text{chr}_i)/\sum_j \text{fitness}(\text{chr}_j)$. Selected chromosomes are crossed over with low (fixed) probability and then the value of each bit of the resulting chromosomes is tweaked with a certain (fixed) mutation probability.

## 5. Application

In this section, we give an example application of the method described in the previous section. We first introduce the example.

**Example 1.** Let $V_0$, $V_1$ and $V_2$ be three categorical variables on ordinal scales. Let

$$L_0 = \{a_1, a_2, a_3, a_4\}$$
$$L_1 = \{b_1, b_2, b_3, b_4, b_5, b_6, b_7, b_8\}$$
$$L_2 = \{c_1, c_2, c_3, c_4, c_5, c_6, c_7\}$$

be the ranges of variables $V_0, V_1$ and $V_2$, respectively. Let $<_{L_0}$, $<_{L_1}$ and $<_{L_2}$ be order relations defined on $L_0$, $L_1$ and $L_2$ according to the position of categories in these sets (e.g. $a_i <_{L_0} a_j$ if $i < j$). Then, consider the set of records

$$R = \{r_1, \ldots, r_{14}\}$$

in Table 1. A graphical representation of these records is given in Fig. 1. This representation shows that $V_0$ is a monotonically increasing function of $V_1$ and $V_2$.

To apply genetic algorithms to Example 1, we have defined a population of 50 chromosomes each consisting of $(|D(V_0)| + 1) + (|D(V_1)| + 1) + (|D(V_2)| + 1) = 5 + 9 + 8 = 22$ integers. Binary translation uses 10 bits for each integer, which yields 220-bit chromosomes.

Table 1
Records used for building a linear model

| Objects | $V_0$ | $V_1$ | $V_2$ |
|---|---|---|---|
| $o_1$ | $a_1$ | $b_1$ | $c_1$ |
| $o_2$ | $a_1$ | $b_2$ | $c_2$ |
| $o_3$ | $a_1$ | $b_3$ | $c_2$ |
| $o_4$ | $a_2$ | $b_3$ | $c_3$ |
| $o_5$ | $a_2$ | $b_4$ | $c_3$ |
| $o_6$ | $a_2$ | $b_4$ | $c_4$ |
| $o_7$ | $a_3$ | $b_5$ | $c_3$ |
| $o_8$ | $a_3$ | $b_5$ | $c_4$ |
| $o_9$ | $a_3$ | $b_6$ | $c_5$ |
| $o_{10}$ | $a_3$ | $b_7$ | $c_5$ |
| $o_{11}$ | $a_3$ | $b_7$ | $c_6$ |
| $o_{12}$ | $a_4$ | $b_8$ | $c_6$ |
| $o_{13}$ | $a_4$ | $b_7$ | $c_7$ |
| $o_{14}$ | $a_4$ | $b_8$ | $c_7$ |

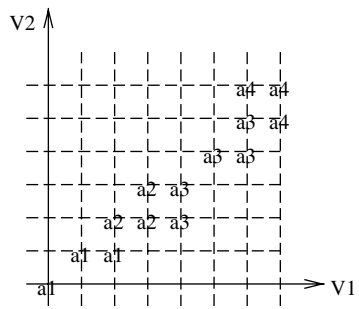


Fig. 1. Graphical representation of the records in Table 1.

The algorithm described in Section 4.1 has been applied and 10 000 iterations have been computed. The best solution after these iterations leads to a fitness equal to:

$$\text{fitness}(\text{best\_chromosome}) = 1975.726$$

This corresponds to a distance (using Expression (3)) equal to $5.06 \times 10^{-4}$. The best solution found consists of the following functions $f_0$, $f_1$ and $f_2$:

$$f_0(a_1) = 0.23, \quad f_0(a_2) = 0.39, \quad f_0(a_3) = 0.66, \quad f_0(a_4) = 0.90$$
$$f_1(b_1) = 0.06, \quad f_1(b_2) = 0.26, \quad f_1(b_3) = 0.39, \quad f_1(b_4) = 0.45$$
$$f_1(b_5) = 0.58, \quad f_1(b_6) = 0.65, \quad f_1(b_7) = 0.78, \quad f_1(b_8) = 0.94$$
$$f_2(c_1) = 0.21, \quad f_2(c_2) = 0.21, \quad f_2(c_3) = 0.24, \quad f_2(c_4) = 0.39$$
$$f_2(c_5) = 0.58, \quad f_2(c_6) = 0.71, \quad f_2(c_7) = 0.84$$

Using these functions $f_0$, $f_1$ and $f_2$, the best model found is:

$$\hat{V}_0(O) = 1.30175 \cdot V_1(O) + 0.67221 \cdot V_2(O)$$

## 6. Conclusions and future work

We have introduced in this paper a method for building models for categorical data in ordinal domains. The method does not assume any continuous variable underlying the ordinal ones. Instead, optimal mapping functions from each categorical domain into the unit interval are found and then a linear model in the latter domain is built. The search of the optimal mapping functions is performed using genetic algorithms. The method has been applied to a toy example to show the feasibility of the approach.

The complexity of the proposed solution is $O(K \cdot N \cdot N \cdot M)$ for each iteration, $N$ being the number of variables, $M$ the number of objects and $K$ the number of chromosomes in each population. By avoiding the repetition of some computations, the complexity can be reduced to $O(K \cdot N \cdot M)$ assuming that $M > N$. Our current implementation performs 17 iterations per second on a desktop PC (using a Java implementation running under a Linux OS).

The genetic algorithm used in Section 5 is rather simple. As the number of objects in Example 1 is pretty small, a large number of iterations is still possible. For a higher number of objects, the operators used and the coding of the solutions could be improved to speed up the convergence of the genetic search. Our current implementation displays an oscillating behaviour of the best fitness in each iteration.

The method proposed in this paper is the ordinal counterpart of numerical linear least-squares regression described in Section 3.1. However, it should be noted that our method can also be regarded as the counterpart of numerical non-linear models such as those described in Section 3.2. This is so because, in the case of ordinal variables, the mapping from the ordinal scale into the continuous one is not fixed a priori, and when fixed, the mapping shapes the space (this has an effect similar to using log or exp transformations on numerical variables).

Future work includes the application of the technique proposed in this paper to a real application and the extension of this methodology to other types of models (e.g. any non-linear model).

## Acknowledgements

This work was partially supported by the European Community under contract "CASC" IST-2000-25069 and by MCyT and FEDER fund under the project "STREAMOBILE" (TIC2001-0633-C03-01/02) is acknowledged.

## References

[1] H.-H. Bock, E. Diday, Analysis of Symbolic Data, Springer, 2000.
[2] J. Domingo-Ferrer, V. Torra, Disclosure methods and information loss for microdata, in: L. Zayatz, J. Lane, P. Doyle (Eds.), Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, North-Holland, Amsterdam, 2001, pp. 113–134.
[3] E.R. Gerber, The privacy context of survey response: an ethnographic account, in: L. Zayatz, J. Lane, P. Doyle (Eds.), Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, North-Holland, Amsterdam, 2001, pp. 371–394.
[4] D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, New York, 1989.
[5] F. Herrera, E. Herrera-Viedma (Eds.), Special Issue on Computing with Words: Foundations and Applications, International Journal of Uncertainty Fuzziness and Knowledge Based Systems, vol. 9, Supplement, 2001.
[6] G. Klir, B. Yuan, Fuzzy Sets and Fuzzy Logic: Theory and Applications, Prentice-Hall, London, 1995.
[7] P. McCullagh, Regression models for ordinal data (with discussion), Journal of the Royal Statistical Society—Series B 42 (1980) 109–142.
[8] S.W. Menard, Applied Logistic Regression Analysis, Sage Publications, 2001.
[9] Z. Michalewicz, Genetic Algorithms + Data Structures = Evolution Programs, Springer, Berlin, 1996.
[10] O. Nelles, Nonlinear System Identification, Springer, Berlin, 2001.