Regression Analysis for Interval-Valued Data

L. Billard[1] and E. Diday[2]

[1] Department Statistics, University of Georgia,
Athens, GA 30602 USA
(e-mail: lynne@stat.uga.edu)
[2] CEREMADE, Universite de Paris 9 Dauphine
75775 PARIS Cedex 16 France
(e-mail: diday@ceremade.dauphine.fr)

Abstract. When observations in large data sets are aggregated into smaller more manageable data sizes, the resulting classifications of observations invariably involve symbolic data. In this paper, covariance and correlation functions are introduced for interval-valued symbolic data. These and their associated terms are then used to fit linear regression models to such data. The methods are illustrated with an example from cardiology.

1 Introduction

Today, we are often faced with the need to make meaningful statistical analyses (such as principal components, discriminant analysis, regression, etc.) on huge data sets whose very size makes the standard form of analysis very difficult to implement. These difficulties may typically exist from a computational viewpoint only even if theoretically there may be no mathematical limitations to the implementation of the relevant statistical analysis. Therefore, before embarking upon a specific analysis, it becomes necessary to classify or to reorganize the data into summary-type classifications or classes, where the number of classes is very much smaller than the number of single individuals in the original data set.

For example, suppose an investigation involves medical records with variables such as pulse rate, blood pressure, disease, etc., as well as identifying variables such as place of residence (Paris, Lyon, London, Brussels,...), age, gender, occupation, etc., for a very large number of individuals. Then, one particular aggregation of the data is to classify individuals by residence (or, by age-gender, occupation, etc.). In another situation, in which data are available for several cities (or regions, or etc.) but already classified by occupation, it may be desired to merge and/or to compare the data for each city whilst still retaining the identifying classification of "occupation". In a different direction, it may be of interest to describe and analyse underlying concepts such as illnesses, species, unemployment, etc. It may be desired to study the issue of whether or not the data contain individuals who might be classified

H. A. L. Kiers et al. (eds.), *Data Analysis, Classification, and Related Methods*

according to some preassigned concept of what the analyst might be seeking. Thus, for example, the starting point could be a query relating to the presence or otherwise of certain diseases.

In these (and similarly related) examples, the resulting data set, after the classification process has been implemented, will almost invariably contain symbolic data rather than classical data values, at least on some (but more probably on all) of the variables describing each observation in the classified data set. Indeed, symbolic data methods may have been an integral part of the classification procedure itself. By symbolic data, we mean that rather than a specific x_j value, an observed value for Y_j can be multi-valued, e.g., $\xi_j = \{1, 4, 7\}$ or {blue, green}, it may be interval-valued, e.g., $\xi_j = [10, 20)$ or $\xi_j = (\geq 0)$; or it may be modal-valued, e.g., $\xi_j = \{1$ with probability .1, 0 with probability .9$\}$, and so on. In addition, there may be rules that have to hold for data integrity, such as the need to maintain underlying information or background knowledge, etc. For example, levels and frequency of cancer (say) treatments must satisfy rules governing the presence of cancer. For a detailed description of symbolic data, see, e.g., Bock and Diday (2000).

Let us suppose we have $(p+1)$ variables Y and X_j, $j = 1, \ldots, p$, with Y being a dependent variable and $\{X_j, j = 1, \ldots, p\}$ being p independent predictor variables, related to Y according to the relation $Y = f(X_1, \ldots, X_p)$. In particular, let us focus attention on a standard linear regression relationship

$$Y = \boldsymbol{X}'\boldsymbol{\beta} + e \tag{1}$$

where $\boldsymbol{X}' = (1, X_1, \ldots, X_p)$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)'$ and $e \sim (0, \sigma^2)$. When the variables X_j take specific values x_j, $j = 0, \ldots, p$, as in classical data, then the appropriate regression analysis for (1) (and indeed for $Y = f(X_1, \ldots, X_p)$ in general) has been very well studied. In this paper, we want to fit regression models of the type (1) to so-called symbolic data. We study herein interval-valued symbolic data; other forms of symbolic data can be handled with appropriate adjustments. This is presented in Section 3 with an example in Section 4. First in Section 2, we look at covariance and correlation functions for interval-valued data.

2 Covariance and correlation functions

Bertrand and Goupil (1999) developed formulae for calculating the univariate empirical frequency distribution, the relative frequency distribution (or equivalently the frequency histogram and hence the empirical distribution function), along with the symbolic empirical mean and variance for interval-valued symbolic data which must also satisfy given logical dependency rules $\{\nu\}$. We extend those ideas to obtain basic descriptive statistics for the two-dimensional variable (Y_1, Y_2), say.

Suppose $u \in E$ is the set of m symbolic objects with observations $Y(u) = \{Y_1(u), Y_2(u)\}$, $u = 1, \ldots, m$. Suppose $Y(u)$ takes specific values on the

rectangle $Z(u) = Y_1(u) \times Y_2(u) = \{\xi_1^u, \xi_2^u\} = ([a_{1u}, b_{1u}], [a_{2u}, b_{2u}])$. Analogously to the univariate case, we assume the individual description vectors $x \in vir(d_u)$ are each uniformly distributed over the rectangle Z(u) where $vir(d_u)$ is the virtual description of x defined as the set of all individual descriptions vectors x which satisfy the set of rules $\{\nu\}$; see Bertrand and Goupil (1999) and Billard and Diday (2000). Then, we can define the empirical joint density function for (Y_1, Y_2) as

$$f(\xi_1, \xi_2) = \frac{1}{m} \sum_{u \in E} \frac{I_u(\xi_1, \xi_2)}{||Z(u)||} \tag{2}$$

where $I_u(\xi_1, \xi_2)$ is the indicator function that (ξ_1, ξ_2) is or is not in the rectangle $Z(u)$ and where $||Z(u)||$ is the area of this rectangle. Note also that the summation in (2) is only over values for which the logical rules hold.

The symbolic empirical covariance between Y_1 and Y_2 is derived according to

$$\begin{aligned} &Cov(Y_1, Y_2) \equiv S_{12} \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\xi_1 - \bar{Y}_1)(\xi_2 - \bar{Y}_2) f(\xi_1, \xi_2) d\xi_1 d\delta_2, \end{aligned} \tag{3}$$

and, substituting from (2) and recalling that $\xi_1 = [a_{1u}, b_{1u}]$ and $\xi_2 = [a_{2u}, b_{2u}]$, we have

$$\begin{aligned} Cov(Y_1, Y_2) &= \frac{1}{m} \sum_{u \in E} \frac{1}{(b_{1u} - a_{1u})(b_{2u} - a_{2u})} \int_{a_{1u}}^{b_{1u}} \int_{a_{2u}}^{b_{2u}} \delta_1 \delta_2 d\delta_1 d\delta_2 - \bar{Y}_1 \bar{Y}_2 \\ &= \frac{1}{m} \sum_{u \in E} \frac{1}{(b_{1u} - a_{1u})(b_{2u} - a_{2u})} \int_{a_{1u}}^{1u} \delta_1 d\delta_1 \int_{a_{2u}}^{b_{2u}} \delta_2 d\delta_2 - \bar{Y}_1 \bar{Y}_2 \\ &= \frac{1}{m} \sum_{u \in E} \frac{1}{(b_{1u} - a_{1u})(b_{2u} - a_{2u})} \left[\frac{1}{2}\delta_1\right]_{a_{1u}}^{b_{1i}} \left[\frac{1}{2}\delta_2\right]_{a_{2u}}^{b_{2u}} - \bar{Y}_1 \bar{Y}_2 \\ &= \frac{1}{4m} \sum_{u \in E} \frac{(b_{1u}^2 - a_{1u}^2)(b_{2u}^2 - a_{2u}^2)}{(b_{1u} - a_{1u})(b_{2u} - a_{2u})} - \bar{Y}_1 \bar{Y}_2. \end{aligned}$$

Hence, the empirical symbolic covariance function is

$$Cov(Y_1, Y_2) = \frac{1}{4m} \sum_{u \in E} (b_{1u} + a_{1u})(b_{2u} + a_{2u}) - \frac{1}{4m^2} \left[\sum_{u \in E} (b_{1u} + a_{1u})\right] \left[\sum_{u \in E} b_{2u} + a_{2u})\right] \tag{4}$$

where the symbolic empirical mean of Y, is, from Bertrand and Goupil (1999),

$$\bar{Y} = \frac{1}{2m} \sum_{u \in E} (b_u + a_u). \tag{5}$$

The symbolic empirical variance of Y is

$$S^2 = \frac{1}{4m} \sum_{u \in E} (b_u + a_u)^2 - \frac{1}{4m^2} \left[\sum_{u \in E} (b_u + a_u)\right]^2. \tag{6}$$

Hence, we can define the symbolic empirical correlation function between two variables Y_1 and Y_2, denoted by $r(Y_1, Y_2)$, as

$$r(Y_1, Y_2) = S_{12}/\sqrt{S_1^2 S_2^2}. \tag{7}$$

3 Linear regression model

Let us take the multiple linear regression model (1). We know that the least squares estimator of the regression coefficient $\boldsymbol{\beta}$ is, for data $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y}). \tag{8}$$

Suppose for simplicity we take $p = 1$. Then, from standard classical theory we have:

$$\hat{\beta}_1 = \frac{Cov(Y, X)}{S_X^2} = r(Y, X)(S_Y/S_X), \tag{9}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}. \tag{10}$$

When the data are symbolic data we can use the same ideas but with the sample means, variances and correlation function terms in (9) and (10) being replaced by their symbolic counterparts. Likewise, we can calculate $\hat{\boldsymbol{\beta}}$ for general p using equation (8).

Notice that since the data are interval-valued and since it is assumed that possible values are uniformly distributed across these intervals, then the formulae for the symbolic means, variances, covariance and correlation funcitons, and hence the regression line fits, correspond to the same values obtained by applying classical methods to the midpoints of the intervals. This would not be the case were the intervals not uniformly distributed or were the data modal or categorical values more generally.

4 Example

Suppose we have the record of the pulse rate Y and the systolic blood pressure X_1 for each of eleven patients (taken from Raju (1997)) and shown in Table 1. Let us suppose there is a logical rule that says the diastolic blood pressure (X_2) must be less than the systolic blood pressure, i.e., $X_2 \leq X_1$. Then, the $u = 7$ observation contradicts this rule, i.e., $|vir(d_u)| = 0$, for $u = 7$. Therefore, in the summation in (2), and hence also in subsequent calculations, terms corresponding to this observation are omitted. Let us find the regression equation

$$Y = \beta_0 + \beta_1 X_1. \tag{11}$$

Table 1 - Data

u	Y Pulse Rate	X_1 Systolic Pressure	X_2 Diastolic Pressure
1	44, 68	90, 100	50, 70
2	60, 72	90, 130	70, 90
3	56, 90	140, 180	90, 100
4	70, 112	110, 142	80, 108
5	54, 72	90, 100	50, 70
6	70, 100	130, 160	80, 110
7	63, 75	60, 100	140, 150
8	72, 100	130, 160	76, 90
9	76, 98	110, 190	70, 110
10	86, 96	138, 180	90, 110
11	86, 100	110, 150	78, 100

First, we calculate the symbolic statistics: $\bar{Y} = 79.1$, $\bar{X}_1 = 131.3$, $S_Y^2 = 162.29$, $S_{X_1}^2 = 495.41$, $Cov(Y, X_1) = 194.170$, and $r(Y, X_1) = 0.685$. Hence, we can calculate $\hat{\beta}_1 = 0.392$, and $\hat{\beta}_0 = 27.639$. Therefore, the regression equation (11) becomes

$$\text{Pulse Rate} = 27.639 + (.392)\ \text{Systolic Pressure}. \tag{12}$$

Suppose now we wanted to predict the pulse rate when the systolic pressure is in the interval $(118, 126)$, say. From (12), we have $Y_{118} = 27.639 + (0.392)(118) = 73.887$ and $Y_{126} = 27.639 + (0.392)(126) = 77.023$. I.e., when the systolic blood pressure is in the interval $(118, 126)$, the predicted pulse rate would be $\xi = (73.89, 77.02)$.

In contrast, had we fitted the regression line (11) through the lower points $\{a_{ju}\}$ only, we would obtain the equation

$$\text{Pulse Rate } (a) = 29.664 + (0.330)\ \text{Systolic Pressure } (a); \tag{13}$$

and likewise by fitting the model (11) through the upper points $\{b_{ju}\}$ only, we obtain the relation

$$\text{Pulse rate } (b) = 45.070 + (0.308)\ \text{Systolic Pressure } (b). \tag{14}$$

Were we to use these relationships to obtain the predicted pulse rate when the systolic pressure is (118,126), we would have the predictions, respectively, Pulse Rate $(a) = (68.60, 71.24)$, and Pulse Rate $(b) = (81.41, 83.88)$. The two regression equations (13) and (14) are standard regression fits as are the corresponding predictions for the systolic blood pressure in the interval (118,126). If we only had these regression fits, it is not unreasonable that we might place greater confidence in the lower level prediction calculation of 68.60 [from pulse

rate (a)], and in the higher level prediction calculation of 83.88 [from pulse rate (b)], giving a prediction interval of (68.60, 83.88). When this is compared with the symbolic prediction interval of (73.89, 77.02), obtained from the symbolic regression equation (12), it is clear that the symbolic analysis gives a tighter fit.

Finally, let us fit both systolic blood pressure (X_1) and diastolic blood pressure (X_2) as predictor variables for pulse rate (Y) i.e., $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. Then, we can show that, for the data of Table 1,

$$X'X = \begin{pmatrix} 10 & 1313 & 846 \\ 1313 & 177351 & 113659 \\ 846 & 113659 & 73396 \end{pmatrix}, \quad X'Y = \begin{pmatrix} 791 \\ 105800 \\ 68329 \end{pmatrix}$$

where the entries in each matrix are the symbolic counterparts of the symbolic sums and crossproducts as given in Section 2. Hence, we have $\hat{\beta}_0 = 20.703$, $\hat{\beta}_1 = -0.040$, and $\hat{\beta}_2 = 0.805$. Therefore, the regression equation becomes

$$\text{Pulse Rate} = 20.703 - (0.040)\ \text{Syst. Press.} + (0.805)\ \text{Diast. Press.} \quad (15)$$

Prediction and other uses of the regression equation (15) then follow in the usual ways.

5 Conclusion

Regression models for other types of symbolic data can be fitted analogously with appropriate adjustments. Thus, for example, if the data were modal, then each rectanglar region $Z(u)$ (or, more generally, a hypercube) would have weights corresponding to the probabilities of the modal data lying in that rectangle. Then, the empirical frequency joint density function $f(\xi_1, \xi_2)$ in (3) would take the necessary form reflecting these weights or probabilities (instead of the uniform rectangular form used above). The subsequent results therefore lead to weighted regression analogues. Logistic regression and generalised linear regression models more broadly defined can likewise be handled.

References

BERTRAND, P. and GOUPIL, F. (1999). Descriptive Statistics for Symbolic Data. In: H.-H. Bock and E. Diday (Eds). *Symbolic Official Data Analysis.* Springer, 103-124.

BILLARD, L. and DIDAY, E. (2000). From the Statistics of Data to a Statistics of Knowledge: Symbolic Data Analysis. In preparation.

BOCK, H. -H. and DIDAY, E. (eds.) (2000). *Symbolic Official Data Analysis.* Springer.

RAJU, S. R. K. (1997). Symbolic Data Analysis in Cardiology. In: E. Diday and K. C. Gowda (Eds). *Symbolic Data Analysis and Its Applications.* CEREMADE, Paris, 245-249.