A Clusterwise Center and Range Regression Model for Interval-Valued Data

Francisco de A. T. de Carvalho, Gilbert Saporta, Danilo N. Queiroz

A Clusterwise Center and Range Regression Model for Interval-Valued Data

Francisco de A. T. de Carvalho[1], Gilbert Saporta[2], and Danilo N. Queiroz[1]

[1] Centro de Informática - CIn/UFPE
Av. Prof. Luiz Freire, s/n - Cidade Universitária, CEP 50740-540, Recife-PE, Brazil {*fatc,dnq*}*@cin.ufpe.br*
[2] Chaire de statistique appliquée & CEDRIC, CNAM
292 rue Saint Martin, Paris, France, *gilbert.saporta@cnam.fr*

Abstract. This paper aims to adapt clusterwise regression to interval-valued data. The proposed approach combines the dynamic clustering algorithm with the center and range regression method for interval-valued data in order to identify both the partition of the data and the relevant regression models, one for each cluster. Experiments with a car interval-valued data set show the usefulness of combining both approaches.

Keywords: clusterwise regression, interval-valued data, symbolic data analysis

1 Introduction

There is a large amount of publications on symbolic interval-valued data (see Billard and Diday (2007)). Symbolic interval-valued data occur in two contexts: either when one has uncertainty on individual values, or when one has variation like eg in medical data such as blood pressure, pulse rate observed on a daily time period. We will consider here only the second case.

Several methods have been proposed to deal with the case where the response y as well as the predictors are interval-valued variables. We will use the centre and range method proposed by Lima Neto and De Carvalho (2008). Assuming that data are homogeneous (ie there is only one regression model for the whole data set) can be misleading. Clusterwise regression has been proposed long ago, as a way to identify both the partition of the data and the relevant regression models, one for each class. Clusterwise regression may be viewed as a particular mixture or latent class model, or from a data analytic perspective as a combination of cluster and regression analysis.

In this paper we adapt clusterwise regression to interval-valued data. The paper is organized as follows. Section 2.1 presents approaches for interval data regression, section 2.2 is a short presentation of clusterwise regression. Section 3 presents how clusterwise regression is extended to interval data. Section 4 presents experiments with a car interval-valued data set in order to show the usefulness of combining both approaches. Finally, section 5 gives concluding remarks.

2 A Brief Overview of Regression for Interval-Valued Data and Clusterwise Linear Regression

2.1 Regression for Interval Data

Billard and Diday (2000) considered the center method where one fits a regression model to the mid-points of the intervals. They predict the bounds of y by applying the model for the centers to the upper bounds of the predictors (resp. the lower bounds). The same model is thus applied to predict the centers and the (upper and lower) bounds. The MinMax method (Billard and Diday (2002)) consists in fitting two different regressions, one for all upper bounds, the other for all lower bounds.

Recently Lima Neto and de Carvalho (2008) presented the *"center and range method"*: in short, this method consists of fitting two linear models, one for the centers of the intervals, another one for the range. The prediction for a new example is given by the prediction of the center $\pm$ the half of the predicted range. In their paper, Lima Neto and de Carvalho (2008) proved with extensive simulations the superiority of the last method compared to the centre and the MinMax method, and it is why we will use it in the following.

2.2 Clusterwise Linear Regression

Clusterwise linear regression is a useful technique when heterogeneity is present in the data. It is a mix of cluster analysis and regression where clusters are obtained in a supervised way in order that for each cluster we have the "best" regression model.

This "local" regression model may also be viewed as a particular mixture model (Wayne et al 1988 and Hennig 2000) who used maximum likelihood estimation. Clusterwise linear regression has been also analyzed in a fuzzy framework (D'Urso and Santoro (2006)). We focus here on least squares techniques. In the basic model the number of clusters is supposed to be known.

Let y be a response variable and $\mathbf{x}$ a p-dimensional vector of regressors. From an algorithmic point of view the aim is to find simultaneously an optimal partition of the data in K clusters, $1 < K < n$ and K regression vectors $\boldsymbol{\beta}_{(k)}$ $(1 < k < K)$ one for each cluster such that one maximizes the overall fit or minimize the sum of squared residuals:

$$\sum_{k=1}^{K} \sum_{i \in P_k} (\epsilon_{i(k)})^2$$

where P_k is the k^{th} cluster, $\hat{y}_{i(k)}$ is the prediction of y (assuming $i \in P_k$) and

$$y_i = (\mathbf{x}_i)^T \boldsymbol{\beta}_{(k)} + \epsilon_{i(k)} = \sum_{j=1}^{p} \beta_{j(k)} x_i j + \epsilon_{i(k)} = \hat{y}_{i(k)} + \epsilon_{i(k)}$$

Numerous algorithms have been proposed to solve this problem: some use combinatorial optimisation techniques like Spaeth (1979) who proposes an exchange algorithm. We will use here the special case of k-means clustering which has been proposed by Diday and Simon (1976) and Bock (1989) and belongs to the family of alternated least squares techniques:
Step 1: Starting from an initial partition, one estimates separately a regression model for each cluster.
Step 2: Each observation is moved to the cluster (or model) giving the smallest square residual (i.e, the best prediction). Once all observations have been reclassified, we obtain a new partition.

Step 1 and 2 are then iterated until convergence (i.e, stability of the partition), or when the criterium does not decrease enough. It is necessary to have enough observations in each cluster (Charles (1977)) in order to estimate the regression coefficients by OLS. Like in k-means clustering, it is possible that some clusters become empty and that the final number of clusters may be less than the initial guess K. Choice of K remains difficult: some have advocated for AIC or BIC- like criteria (Plaia 2001)). From a empirical machine learning point of view, K should be chosen by some validation technique (cross-validation, bootstrap. etc.). The existence of many local minima have been stressed by Caporossi and Hansen (2007): this implies to choose wisely the starting partition.

3 Clusterwise regression on interval-valued data

This section presents a clusterwise regression model based on both the dynamic clustering algorithm (Diday and Simon (1976)) and the center and range regression model for interval-valued data (Lima Neto and De Carvalho (2008)).

Let $E = \{1, \ldots, n\}$ be a set of observations that are described by $p + 1$ interval-valued variables $z, w_1, \ldots, w_p$. Each observation $i \in E\,(i = 1, \ldots, n)$ is represented by a vector of intervals $\mathbf{e}_i = (w_{i1}, \ldots, w_{ip}, z_i)$, where $w_{ij} = [w_{ij}^L, w_{ij}^U]\,(j = 1, \ldots, p)$ and $z_i = [z_i^L, z_i^U]$.

Let $\mathbf{y}$ and $\mathbf{x}_j$ $(j = 1, \ldots, p)$ be, respectively, quantitative bi-variate variables that assume as their values the midpoints and half ranges of the interval assumed by the interval-valued variables z and w_j. Thus, each observation $i \in E\,(i = 1, \ldots, n)$ is also represented as a vector of bi-variate quantitative vectors $\mathbf{t}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{ip}, \mathbf{y}_i)$, with

$$\mathbf{x}_{ij} = \begin{pmatrix} x_{ij}^c \\ x_{ij}^r \end{pmatrix}\ (j = 1, \ldots, p) \text{ and } \mathbf{y}_i = \begin{pmatrix} y_i^c \\ y_i^r \end{pmatrix}$$

where $x_{ij}^c = (w_{ij}^L + w_{ij}^U)/2$, $x_{ij}^r = (w_{ij}^U - w_{ij}^L)/2$, $y_i^c = (z_i^L + z_i^U)/2$ and $y_i^r = (z_i^U - z_i^L)/2$.

This clusterwise regression model for interval-valued data looks for a partition of E in K clusters $P_1, \ldots, P_K$, each cluster being represented by a

prototype, such that an adequacy criterion measuring the fit between the clusters and their prototypes are locally minimized. The particularity of this kind of method is that the prototype of each cluster is represented by the hyper-plane given by the linear regression relationship between the dependent variable and the independent predictor variables:

$$\mathbf{y}_{i(k)} = \boldsymbol{\beta}_{0(k)} + \sum_{j=1}^{p} \boldsymbol{\beta}_{j(k)}\, \mathbf{x}_{ij} + \boldsymbol{\epsilon}_{i(k)}\, (\forall i \in P_k) \qquad \text{where} \tag{1}$$

$$\boldsymbol{\beta}_{0(k)} = \begin{pmatrix} \beta^c_{0(k)} \\ \beta^r_{0(k)} \end{pmatrix}, \quad \boldsymbol{\beta}_{j(k)} = \begin{pmatrix} \beta^c_{j(k)} & 0 \\ 0 & \beta^r_{j(k)} \end{pmatrix}\, (j = 1, \ldots, p) \qquad \text{and}$$

$$\boldsymbol{\epsilon}_{i(k)} = \begin{pmatrix} \epsilon^c_{i(k)} \\ \epsilon^r_{i(k)} \end{pmatrix} = \begin{pmatrix} y^c_i - \left(\beta^c_{0(k)} + \sum_{j=1}^{p} \beta^c_{j(k)}\, x^c_{ij}\right) \\ y^r_i - \left(\beta^r_{0(k)} + \sum_{j=1}^{p} \beta^r_{j(k)}\, x^r_{ij}\right) \end{pmatrix}\, (\forall i \in P_k)$$

The adequacy criterion is defined as:

$$J = \sum_{k=1}^{K} \sum_{i \in P_k} (\boldsymbol{\epsilon}_{i(k)})^T \boldsymbol{\epsilon}_{ik} = \sum_{k=1}^{K} \sum_{i \in P_k} \left[(\epsilon^c_{i(k)})^2 + (\epsilon^r_{i(k)})^2 \right] \tag{2}$$

$$= \sum_{k=1}^{K} \sum_{i \in P_k} \left\{ \left[y^c_i - \left(\beta^c_{0(k)} + \sum_{j=1}^{p} \beta^c_{j(k)}\, x^c_{ij} \right) \right]^2 + \left[y^r_i - \left(\beta^r_{0(k)} + \sum_{j=1}^{p} \beta^r_{j(k)}\, x^r_{ij} \right) \right]^2 \right\}$$

This algorithm sets an initial partition and alternates two steps until convergence when the criterion J reaches a local minimum.

3.1 Step 1: definition of the best prototypes

In the first stage, the partition of E in K clusters is fixed.

Proposition 1. *The prototype*

$$\hat{\boldsymbol{y}}_{i(k)} = \begin{pmatrix} \hat{y}^c_{i(k)} \\ \hat{y}^r_{i(k)} \end{pmatrix} = \begin{pmatrix} \hat{\beta}^c_{0(k)} + \sum_{j=1}^{p} \hat{\beta}^c_{j(k)}\, x^c_{ij} \\ \hat{\beta}^r_{0(k)} + \sum_{j=1}^{p} \hat{\beta}^r_{j(k)}\, x^r_{ij} \end{pmatrix}\, (\forall i \in P_k)$$

of cluster $P_k\, (k = 1, \ldots, K)$ *has the least square estimates of the parameters* $\hat{\beta}^c_{j(k)}$ *and* $\hat{\beta}^r_{j(k)}\, (j = 0, 1, \ldots, p)$*, which minimizes the clustering criterion* J*, given by the solution of the system of* $2\,(p+1)$ *equations:*

$$\hat{\boldsymbol{\beta}} = \left(\hat{\beta}^c_{0(k)}, \hat{\beta}^c_{1(k)}, \ldots, \hat{\beta}^c_{p(k)}, \hat{\beta}^r_{0(k)}, \hat{\beta}^r_{1(k)}, \ldots, \hat{\beta}^r_{p(k)} \right)^T = (\mathbf{A})^{-1}\mathbf{b} \tag{3}$$

where $\mathbf{A}$ *is a matrix* $2(p+1) \times 2(p+1)$ *and* $\mathbf{b}$ *is a vector* $2(p+1) \times 1$*, denoted as:*

$$\mathbf{A} = \begin{pmatrix} |P_k| & \sum_{i\in P_k} x_{i1}^c & \cdots & \sum_{i\in P_k} x_{ip}^c & 0 & 0 & \cdots & 0 \\ \sum_{i\in P_k} x_{i1}^c & \sum_{i\in P_k} (x_{i1}^c)^2 & \cdots & \sum_{i\in P_k} x_{ip}^c x_{i1}^c & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{i\in P_k} x_{ip}^c & \sum_{i\in P_k} x_{i1}^c x_{ip}^c & \cdots & \sum_{i\in P_k} (x_{ip}^c)^2 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & |P_k| & \sum_{i\in P_k} x_{i1}^r & \cdots & \sum_{i\in P_k} x_{ip}^r \\ 0 & 0 & \cdots & 0 & \sum_{i\in P_k} x_{i1}^r & \sum_{i\in P_k} (x_{i1}^r)^2 & \cdots & \sum_{i\in P_k} x_{ip}^r x_{i1}^r \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \sum_{i\in P_k} x_{ip}^r & \sum_{i\in P_k} x_{i1}^r x_{ip}^r & \cdots & \sum_{i\in P_k} (x_{ip}^r)^2 \end{pmatrix}$$

$$\mathbf{b} = \left(\sum_{i\in P_k} y_i^c, \sum_{i\in P_k} y_i^c x_{i1}^c, \ldots, \sum_{i\in P_k} y_i^c x_{ip}^c, \sum_{i\in P_k} y_i^r, \sum_{i\in P_k} y_i^r x_{i1}^r, \ldots, \sum_{i\in P_k} y_i^r x_{ip}^r \right)^T$$

3.2 Step 2: definition of the best partition

In this step, the prototypes $\hat{\mathbf{y}}_{i(k)}\,(k = 1, \ldots, K)$ are fixed.

Proposition 2. *The optimal clusters $P_k\,(k = 1, \ldots, K)$, which minimize the criterion J, are obtained according to the following allocation rule:*

$$P_k = \{i \in E : (\boldsymbol{\epsilon}_{i(k)})^T \boldsymbol{\epsilon}_{i(k)} \leq (\boldsymbol{\epsilon}_{i(h)})^T \boldsymbol{\epsilon}_{i(h)},\ \forall h \neq k\,(h = 1, \ldots, K)\} \tag{4}$$

Given a new observation $\mathbf{e} = (w_1, \ldots, w_p, z)$ described by the vector of bivariate quantitative vectors $\mathbf{t} = (\mathbf{x}_1, \ldots, \mathbf{x}_p, \mathbf{y})$, the interval $z = [z^L, z^U]$ is predicted from the estimated bivariate vector $\hat{\mathbf{y}}_{(k)} = (\hat{y}_{(k)}^c, \hat{y}_{(k)}^r)\,(k = 1, \ldots, K)$, as follows

$$\hat{z}_{(k)}^L = \hat{y}_{(k)}^c - \hat{y}_{(k)}^r \text{ and } \hat{z}_{(k)}^U = \hat{y}_{(k)}^c + \hat{y}_{(k)}^r$$

where $\hat{y}_{(k)}^c = \hat{\beta}_{0(k)}^c + \sum_{j=1}^p \hat{\beta}_{j(k)}^c\, x_j^c$ and $\hat{y}_{(k)}^r = \hat{\beta}_{0(k)}^r + \sum_{j=1}^p \hat{\beta}_{j(k)}^r\, x_j^r$.

"Goodness-of-fit measures" (determination coefficients) for these clusterwise regression models are computed, for k=1,...,K, as:

$$R_{c(k)}^2 = \frac{\sum_{i\in P_k} \left(\hat{y}_{i(k)}^c - \bar{y}_{c(k)}\right)^2}{\sum_{i\in P_k} \left(y_i^c - \bar{y}_{c(k)}\right)^2}\ ;\ R_{r(k)}^2 = \frac{\sum_{i\in P_k} \left(\hat{y}_{i(k)}^r - \bar{y}_{r(k)}\right)^2}{\sum_{i\in P_k} \left(y_i^r - \bar{y}_{r(k)}\right)^2} \tag{5}$$

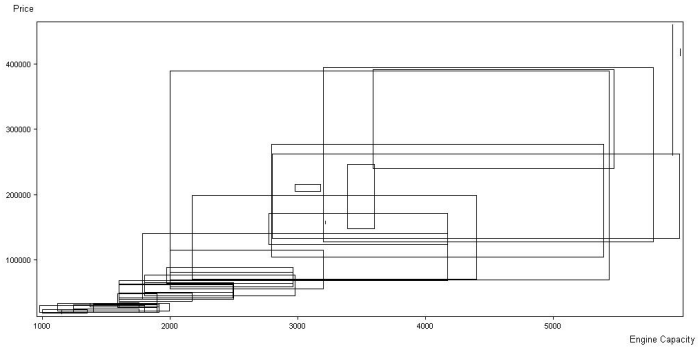

Fig. 1. The car interval-valued data set.

where $\bar{y}_{c(k)} = \sum_{i \in P_k} y_i^c / n$, $\bar{y}_{r(k)} = \sum_{i \in P_k} y_i^r / n$ and $R^2_{c(k)}$, $R^2_{r(k)}$ are, respectively, the determination coefficient of "center" and "range" models.

Other measures, in order to obtain the performance assessment of these linear regression models, are the lower ($RMSE_L$) and the upper ($RMSE_U$) boundaries root-mean-square error. They are computed as

$$RMSE_L = \sqrt{\frac{\sum_{i=1}^{n} \left(z_i^L - \hat{z}_i^L\right)^2}{n}} \; ; \; RMSE_U = \sqrt{\frac{\sum_{i=1}^{n} \left(z_i^U - \hat{z}_i^U\right)^2}{n}} \tag{6}$$

4 Application: a car interval-valued data set

The car data set[1] (Figure 1) consists of a set of 33 car models described by 2 interval-valued variables: price y and engine capacity x_1. The aim is to predict the interval values of y (the dependent variable) from x_1 through linear regression models. In this application, the 2 interval-valued variables – Price and Engine Capacity –, have been considered for clustering purposes. The clusterwise regression algorithm has been performed on this data set in order to obtain a partition in $K = \{1, 2, 3\}$ clusters. For a fixed number of clusters K, the clustering algorithm is run 100 times and the best result according to the adequacy criterion is selected.

Table 1 presents the regression equations fitted over the car interval-valued data set. Table 2 gives the determination coefficients for the 1-cluster, 2-cluster and 3-cluster partitions.

In order to obtain a better preditive model, the estimates of the K regression models given by the K-cluster partition ($K = 1, 2, 3$), obtained with this algorithm, were combined according to the "stacked regressions" approach. According to Breiman (1996), this approach uses cross validation data and

[1] This data set is available with the SODAS software at `http://www.info.fundp.ac.be/asso/index.html`.

Table 1. Fitted regression equations over the whole car interval-valued data set

$K-partition$	cluster k	"Center Model"	"Range Model"
1	1	$\hat{y}^c_{(1)} = -98840.9 + 79.2\,x^c_1$	$\hat{y}^r_{(1)} = -341.4 + 60.9\,x^r_1$
2	1	$\hat{y}^c_{(1)} = -63462.2 + 59.6\,x^c_1$	$\hat{y}^r_{(1)} = -4560.1 + 47.1\,x^r_1$
	2	$\hat{y}^c_{(2)} = -22836.5 + 68.8\,x^c_1$	$\hat{y}^r_{(2)} = 34563.6 + 68.6\,x^r_1$
3	1	$\hat{y}^c_{(1)} = -77422.1 + 82.0\,x^c_1$	$\hat{y}^r_{(1)} = 2229.7 + 92.2\,x^r_1$
	2	$\hat{y}^c_{(2)} = -58484.1 + 71.1\,x^c_1$	$\hat{y}^r_{(2)} = 101952.9 - 546.7\,x^r_1$
	3	$\hat{y}^c_{(3)} = -73362.1 + 62.0\,x^c_1$	$\hat{y}^r_{(3)} = -9755.9 + 53.2\,x^r_1$

Table 2. Determination coefficients for the fitted regression equations over the whole car interval-valued data set

K-partition	1	2		3		
cluster k	1	1	2	1	2	3
$R^2_{c(k)}$	0.93	0.95	0.91	0.97	0.99	0.98
$R^2_{r(k)}$	0.53	0.79	0.66	0.98	0.98	0.83

least squares under non-negativity constraints for forming linear combinations of different predictors to give improved prediction accuracy.

The car interval-valued data set $\mathcal{L}$ was partitioned into 10 folds $\mathcal{L}_{(j)}$ $(j = 1, \ldots, 10)$ of size as nearly equal as possible. For a fixed number of clusters K, the clustering algorithm is run 100 times on 9 folds $\mathcal{L}^{(j)} = \mathcal{L} - \mathcal{L}_{(j)}$ and the best result according to the adequacy criterion is selected. The K regression models are used to give preditions for the lower and upper boundary of the dependent variable on the $\mathcal{L}^{(j)}$ learning data set. These predictions were combined according to the "stacked regressions" approach to obtain the predictions for the observations belonging to the test data set $\mathcal{L}_{(j)}$. The $RMSE_L$ and $RMSE_U$ measures are computed from the predicted values on the test data sets $\mathcal{L}_{(j)}$ $(j = 1, \ldots, 10)$.

This process is repeated 100 times and it is calculated the average and standard deviation of the $RMSE_L$ and $RMSE_U$ measures (Table 3). Even if the observed mean differences are not statistically significant, we can conclude that 2 regression models given by the 2-cluster partition give the best preditive model through the "stacked regressions" approach.

Table 3. Average Root-mean-square error for the combined estimates of the K regression models

K-partition	1	2	3
$RMSE_L$	96649.28 (13812.49)	90417.42 (13538.22)	94993.75 (11376.24)
$RMSE_U$	143416.6 (17294.02)	135471.4 (17027.49)	137825.9 (14243.29)

5 Concluding Remarks

This paper introduced a suitable clusterwise regression model for interval-valued data. The proposed model combines the dynamic clustering algorithm with the center and range regression model for interval-valued data in order to identify both the partition of the data and the relevant regression models (one for each cluster). Experiments with a car interval-valued data set showed the interest of this approach. Other experiments on medical data sets are in progress.

References

BILLARD, L. and DIDAY, E. (2000): Regression Analysis for Interval-Valued Data. In: H.A.L. Kiers, J.P. Rasson, P.J.F. Groenen and M. Schader (Eds.): *Data Analysis, Classification and Related Methods*. Springer, Berlin, 369–374.

BILLARD, L. and DIDAY, E. (2007): *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley-Interscience, San Francisco.

BILLARD, L. and DIDAY, E. (2002): Symbolic Regression Analysis. In: K. Jajuga, A. Sokolowski and H.-H. Bock (Eds.): *Classification, Clustering, and Data Analysis. Recent Advances and Applications*. Springer, Berlin, 281–288.

BOCK, H.-H. (1989): The equivalence of two extremal problems and its application to the iterative classification of multivariate data. *Lecture* note. Mathematisches Forschungsinstitut Oberwolfach.

BREIMAN, L. (1996): Stacked Regressions. *Machine Learning 24, 49–64*.

CAPAROSSI, G. and HANSEN, P. (2007): Variable Neighborhood Search for Least Squares Clusterwise Regression. *Les Cahiers du GERAD, G 2005-61*. HEC Montréal

CHARLES, C. (1977): Régression typologique et reconnaissance des formes. *Thèse de doctorat* Université Paris IX.

DIDAY, E. and SIMON, J.C. (1976): Clustering analysis. In: K.S. Fu (Eds.): *Digital Pattern Classification*. Springer, Berlin, 47–94.

D'URSO, P. and SANTORO, A. (2006): Fuzzy clusterwise linear regression analysis with symmetrical fuzzy output variable. *Computational Statistics and Data Analysis, 51 (1): 287–313*.

GONZÁLEZ-RODRIGUES, G., BLANCO, A., CORRAL, N. and COLUBI, A. (2007): Least squares estimation of linear regression models for convex compact random sets. *Advances in Data Analysis and Classification 1, 67–81*.

HENNIG, C. (2000): Identifiability of models for clusterwise linear regression. *J. Classification 17 (2), 273–296*.

LIMA NETO, E. A. and DE CARVALHO, F.A.T. (2008): Centre and Range method for fitting a linear regression model to symbolic interval data. *Computational Statistics and Data Analysis, 52 (3): 1500–1515*.

PLAIA, A. (2001): On the number of clusters in clusterwise linear regression. In: *Xth International Symposium on Applied Stochastic Models and Data Analysis, Proceedings, vol. 2*. Compiegne, France, 847–852.

SPAETH, H. (1979): Clusterwise Linear Regression. *Computing 22 (4), 367–373*.

WAYNE, S., DESARBO, W.S. and CRON, W.L. (1988): A maximum likelihood methodology for clusterwise linear regression. *J. Classification 5 (2), 249–282*.