Rescaling Process by Using the Constrained L_1 Norm Method

M. A. K. Al-Mansoob[1]

Abstract :

The constrained L_1 norm method (CL1) was used to rescale ordered variables in a regression problem with a continuous dependent variable. The rescaling process has lead to new scoring systems to the ordered variables that effectively improved the estimation accuracy. The rescaling process may also be adopted for variable selection.

[1]Mathematics Department, Faculty of Science, Sana'a University

1. Introduction

For a data set with a quantitative dependent variable and ordered predictor variables, the following model has often been applied:

$$y_i = \beta_0 + \beta_1 \sum_j x_{ij} + \varepsilon_i, \qquad i = 1,2,...,n \qquad (1)$$

Where y_i is the value of the dependent variable in the ith individual, β_0 and β_1 are parameters to be estimated, x_{ij} is the score of the jth independent variable on the ith individual, so that $\sum_j x_{ij}$ is the total score for the ith individual and ε_i is random error term with mean $E(\varepsilon_i) = 0$ [1,2,3]. Often $\{\varepsilon_i\}$ is assumed to be a sequence of iid $N(0,\sigma^2)$ random variables. Usually the scores are non-negative integers: sometimes the scores are given the values 0,1,2,3,... .

Model (1) might be inappropriate for any of several reasons. If there are outliers, the normality assumption might be grossly violated, and so the usual least square approach is inappropriate. Here we consider minimum absolute deviations and so give less weight to outliers. This provides maximum likelihood estimates if the error term ε_i is assumed to follow a double exponential distribution. Furthermore, the scores might need reallocating subject to certain properties: for our envisaged uses, the scores should be positive, integers and ordered within each independent

variable. Later we will see how such a scoring system can be derived and how the system can be used to reduce the number of independent variables in the estimation procedure. We show how to apply a constrained least absolute deviation algorithm, CL1 in the nomenclature adopted.

2. Rescaling Process

The rescoring or the rescaling process starts by recoding the independent variables in model (1) .Let

$$y_i = \alpha_0 + \sum_{j.k} \alpha_{jk} v_{ijk} + \varepsilon_i \quad , \quad i = 1,2,...,n \qquad (2)$$

Where y_i is the ith value of the dependent variable, α_0 and $\alpha_{j.k}$ for all pairs (i, j) are parameters, and x_{ij} is the jth independent variable on the ith individual.

$$v_{ijk} = \begin{cases} 1 \ if \ x_{ij} = k \\ 0 \ if \ x_{ij} \neq k, \ \ and \ \alpha_{j0} = 0 \ \forall \ j \end{cases}$$

The new design matrix V contains the variables that are the components of each characteristic: these new variables v_{ijk} are of binary form and the arrangement for the ith observation is considered as a vector v_i. The next step is to use a suitable multiple regression process with all the new independent variables. These parameter estimates form the new scoring system. Although collinearity or multicollinearity can arise among the new

independent variables, available algorithms do not require the design matrix to be of full rank. In addition [1,2], the collinearity or the multicollinearity can be adjusted in the definition of the variables themselves: for example, the score of the smallest ordered value of each variable should be set to zero.

2.1 CL1 Formulation

By considering model (2) in matrix notation, the L_1 norm minimizes

$$\left\| \underline{Y} - V\underline{\alpha} \right\|_1 = \sum_i \left| y_i - v_i \underline{\alpha} \right| \qquad (3)$$

The *CL*1 algorithm of Barrodale and Roberts [4,5] minimizes (3) subject to:

$$A\underline{\alpha} = \underline{B} \qquad (4)$$
$$C\underline{\alpha} = \underline{D}$$

Where A and C are matrices of dimensions $l \times m$ and $k \times m$ respectively, $\underline{B}$ and $\underline{D}$ are columns vectors of dimensions l and k, respectively. Matrix C accommodates the fact that the scores are ordered in specific ways that reflect the physical process. Barrodale and Roberts have mentioned that matrices V, A and C are not required to be of full rank [4]. This property is important because of the likely multicollinearity among the independent variables. Barrodale and Roberts [4] reformulated the problem represented by (3) and (4) as a linear program:

$$Z = \min \sum_i \left| e_i^+ + e_i^- \right| = \min (\underline{e}^+ + \underline{e}^-), \; i = 1,2,...,n$$

subject to

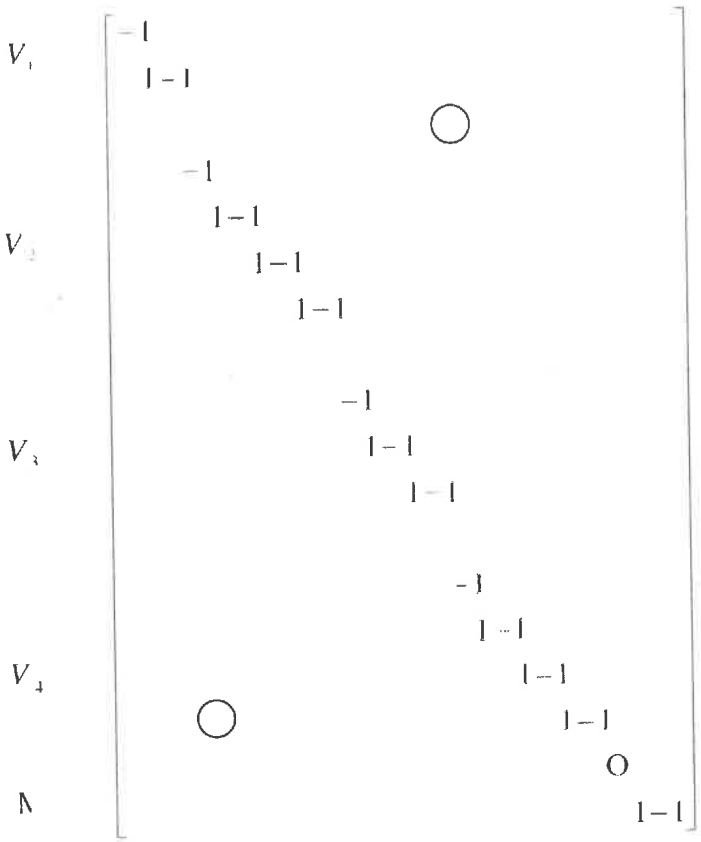
$$\underline{Y} = V(\underline{\hat{\alpha}}^{+} - \underline{\hat{\alpha}}^{-}) + \underline{e}^{+} - \underline{e}^{-}$$
$$\underline{B} = A(\underline{\hat{\alpha}}^{+} - \underline{\hat{\alpha}}^{-})$$
$$\underline{D} = C(\underline{\hat{\alpha}}^{+} - \underline{\hat{\alpha}}^{-}) + \underline{e}^{''+}$$
$$\underline{\hat{\alpha}}^{+}, \underline{\hat{\alpha}}^{-} \geq 0, \underline{e}^{+}, \underline{e}^{-} \geq 0, \underline{e}^{''+} \geq 0$$

Where $\underline{e}^{+}$ and $\underline{e}^{-}$ are $n \times 1$ column vectors and $\underline{e}^{''+}$ is an $m \times 1$ vector. The algorithm uses the two-phase simplex methods (more detailed about these methods can be seen on Barrodale and Roberts and Taha [4,6]). In a particular problem we faced, the parameter estimates had to be ordered as follows:

For matrix V_1 corresponding to 1st variable, $0 \leq \hat{\alpha}_1 \leq \hat{\alpha}_2$, and for the remaining variables

$$V_2 \rightarrow \ 0 \leq \hat{\alpha}_3 \leq \hat{\alpha}_4 \leq \hat{\alpha}_5 \leq \hat{\alpha}_6$$
$$V_3 \rightarrow \ 0 \leq \hat{\alpha}_7 \leq \hat{\alpha}_8 \leq \hat{\alpha}_9$$
$$V_4 \rightarrow \ 0 \leq \hat{\alpha}_{10} \leq \hat{\alpha}_{11} \leq \hat{\alpha}_{12} \leq \hat{\alpha}_{13}$$
M

The matrix C for this structure takes the form:



To use the published algorithm, C is placed under the design matrix V to form a composite matrix.

The imposition of these constraints ensures the positiveness and the order within each of the original variables. By equating A to 0, the equality constraints are unnecessary.

2.2 Application

The gestational age (G.A) of a newborn baby is often estimated by observing some physical characteristics, and some characteristics are shown in Table 1. Dubowitz et al (1970) assigned scores to each characteristic so that the higher the score the greater the maturity where, of course, that these characteristics are continuous [7]. Estimating the gestational age then is often done by regressing the total sum (T.S.) of the characteristics scores against the actual gestational ages [8,9,10,11].

The normality assumption to the error term in model (1) appeared to be violated when the least square method was used for two data sets: the first for 200 Sudanese babies and the second for 144 British babies [11].

3. Results and Discussion

Model (2) is fitted to both the Sudanese and the British data (the final characteristic is omitted from the Sudanese study) by using the constrained L_1 ($CL1$) Algorithm of Barrodale and Roberts (1980) [5]. The parameter estimates (Tables 2) could be used for the new scoring systems. Once the new scoring systems have been constructed, the total sum of the scores of the remaining variables is recalculated in model (1) and β_0 and β_1 are estimated by minimizing the absolute deviations. Those variables in Tables 2 with parameter estimates zero are omitted. While no integer constraint were imposed initially, the parameter estimates in Table 2 are seen

to be integers which is a desirable property given the likely use of the results. The estimates are integers because the data have been transformed into binary form.

These integer-valued estimates led to consider the problem as an exercise in integer programming. In integer programming the problem can be reformulated as follows:

$$Z = \min \left| \underline{e} \right|$$

subject to

$$\underline{Y} - V\,\underline{\hat{\alpha}} = \underline{e}$$

$$C\,\underline{\hat{\alpha}} \le \underline{0}$$

$$\underline{Y}, V, C \ge 0, \underline{\hat{\alpha}} \ge 0 \qquad \text{with integer elements}$$

The Branch and Bound method was applied, but no improvements in computational time achieved. *CL*1 analysis results are summarized and compared in Table 3. The original and the new scoring systems are used to estimate the gestational age of model (1). In *CL*1 the means of the sum absolute deviations (MSAD) are given for comparison: the procedure reduces the values of MSAD for both data sets and there is less variability about the line in the MSAD sense for the British data then for the Sudanese data.

Conclusion

The scores of ordered variables are usually to distinguish the progressing level of the variables. In analysing data with ordered variables, trivial questions may arise "Is the utilized scoring system efficient?' And if not "Are there any possible ways for rescaling or rescoring the variables?". The proposed method for rescaling ordered variables in this paper is mainly a regression procedure. However, if data of the ordered variables that to be rescaled are showing some erroneous values (outliers), using the ordinary least square method is inappropriate. In a previous work, Al-Mansoob (1998) has found that some outliers are present in the two sets of the data; the Sudanese and the British [11]. Hence, it was necessary to seek other regression procedures that give less weight to the outliers effect; L_1 norm method was used here. Since the rescaled values should be positive and ordered within each variable, the algorithm of Barrodale and Roberts (1980), that based on L_1 norm procedure, was found suitable [5]. Applying this algorithm on the two sets of data has produced ordered, positive and unexplainable integer values new scoring systems to some variables and eliminates some others. When the remaining variables have been reallocated with the derived scores, the estimation accuracy has improved. Therefore, the proposed rescoring procedure is worthy to be tested and evaluated.

References

[1] Draper, N., Smith, H. (1981). Applied Regression Analysis. Wiley, 2nd Edn.

[2] Neter, J., Wasserman, W., Kutner, M.H. (1985). Applied Linear Statistical Models. Irwin.

[3] Everitt, B.S. (1994). Statistical Methods for Medical Investigations. Edward Arnold, 2nd Edn.

[4] Barrodale, I., Roberts, F.D.K.(1978). An Efficient Algorithm for Discrete L_1 Linear Approximation with Linear Constraints. SIAM J. Numer. Anal.; Vol. 15, No.3: 603-611.

[5] Barrodale, I., Roberts, F.D.K.(1980). Algorithm 552 Solution for the constrained L_1 Linear Approximation Problem (F4). ACM Transactions on Mathematical Software; Vol.6, No.2: 231-235.

[6] Taha. A.H. (1987). Operations Research: An Introduction. Collier Macmillan. Fourth Edition.

[7] Dubowitz, L.M.S, Dubowitz, V., and Goldberg, C. (1970). Clinical Assessment of Gestational Age in the Newborn. Journal of Paediatrics; 77, No. 1:1-10.

[8] Shu-Zhong, S., Shang-Ping, Q., Hang-Wei, Y., Fang-Ying, Z., Yu-Fang, F, Yi-Ming, Z. (1982). Assessment of gestational age. Chinese Medical Journal; 95: 777-780.

[9] Downham D.Y., Dixon T.J., Elshibly E., Omer M.(1988). How old are babies at birth?. Sudan MJ; 24:1-4.

[10] Elshibly, E.M., Omer, M.I., Downham, D.Y.(1985). Assessment of gestational age in Sudanese newborns. Annals of Tropical Paediatrics; 5: 76-68.

[11] Al-Mansoob, M.A.K (1998). New physical scoring systems for estimating gestational age in Sudanese and the British babies. Journal of Decision and Mathematika Sciences (J-DAMS); 3(1-3): 45-54, India.

Table 1: Dubowitz, Dubowitz and Goldberg Original Scoring System

Variables		Original Scores				
Oedema	x_1	0	1	2		
Skin Texture	x_2	0	1	2	3	4
Skin Color	x_3	0	1	2	3	
Skin Opacity	x_4	0	1	2	3	4
Lanugo (Over Back)	x_5	0	1	2	3	4
Plantar Creases	x_6	0	1	2	3	4
Nipple Formation	x_7	0	1	2	3	
Breast Size	x_8	0	1	2	3	
Ear Formation	x_9	0	1	2	3	
Ear Firmness	x_{10}	0	1	2	3	
Genitalia	x_{11}	0	1	2		
Skull Hardness	x_{12}	0	1	2	3	4

Tables 2: The New Scoring Systems under CL1 Approximations

2.1 The Sudanese Babies

Variables		New Scores				
Oedema	x_1	0	21	23		
Skin Texture	x_2	0	5	11	11	11
Skin Color	x_3	0	13	14	14	
Skin Opacity	x_4	0	0	1	1	1
Lanugo (Over Back)	x_5	0	0	0	0	0
Plantar Creases	x_6	0	0	0	0	1
Nipple Formation	x_7	0	0	0	0	
Breast Size	x_8	0	13	23	26	
Ear Formation	x_9	0	0	0	1	
Ear Firmness	x_{10}	0	0	31	33	
Genitalia	x_{11}	0	1	2		

2.1 The British Babies

Variables		New Scores				
Oedema	x_1	0	0	0		
Skin Texture	x_2	0	0	0	3	18
Skin Color	x_3	0	0	9	11	
Skin Opacity	x_4	0	0	6	6	7
Lanugo (Over Back)	x_5	0	0	2	4	4
Plantar Creases	x_6	0	0	0	1	1
Nipple Formation	x_7	0	0	0	0	
Breast Size	x_8	0	·	8	8	
Ear Formation	x_9	0	3	3	3	
Ear Firmness	x_{10}	0	0	0	1	
Genitalia	x_{11}	0	0	3		
Skull Hardrness	x_{12}	·	·	21	21	21

Table 3 : Fitting the Data According to Model (1) after Reallocating the Scores for CL1 Analysis

Scoring System Used	Country	No. of Variables	Equation Used to Estimate G.A.	MSAD
Original	Sudan	11	215.57 + 2.21 * T.S.	12.08
Original	Britain	12	248.67 + 1.20 * T.S.	07.19
New	Sudan	9	164.00 +1.00 * T.S.	09.84
New	Britain	10	225.00 +1.00 * T.S.	06.00

عملية إعادة تدرج متغيرات مرتبة باستخدام طريقة الانحرافات المطلقة المقيدة

محمد المنصوب*

خلاصة :

لتقدير متغير كمي متصل تابع من خلال مجموع قيم متغيرات مرتبة مستقلة قد لا يكون نموذج الانحدار الخطي البسيط مناسبا لعدة أسباب أهمها وجود قيم شاذة في البيانات والذي يعني إنتهاء شرط التورخ الطبيعي للبقايا (Residuals) . الأمر الذي هو أحد أهم الشروط الأساسية لاستخدام طريقة المربعات الصغرى أو أقل المربعات في التقدير ولمعالجة مثل هذا الوضع هناك طرق تقدير مختلفة لتقليل تأثير القيم الشاذة اختير منها طريقة أقل الانحرافات المطلقة والمصطلح عليها L_1-Norm (Least Absolute Deviation) أو LAD (Least Absolute Deviation) أو LAV (Least Absolute Variation) أو MAD (Minimum Absolute Deviation) والتي يشيع استخدامها في الكثير من مسائل التحليل العددي

شذوذ المجموع الكلي للمتغيرات المرتبة المستقلة ربما يأتي من أن نظام التدرج في هذه المتغيرات غير مناسبا وبالتالي وإذا كان الأمر كذلك فهل بالإمكان إيجاد أنظمة تدرج أفضل لهذه المتغيرات مع الاحتفاظ بنفس خصائص التدرج كأن تكون القيم موجبة وصحيحة وكذلك مرتبة داخل كل متغير باستخدام طريقة أقل الانحرافات المطلقة المقيدة (Constrained L1-Norm Method) أو اختصارا CL1 أمكن استنتاج أنظمة تدرج جديدة لمتغيرات مستقلة مرتبة من خلال أنظمة التدرج القديمة حيث عند استخدامها -الجديدة - على نفس البيانات أمكن الحصول على نتائج تقدير أكثر دقة

عملية إعادة قيم التدرج بهذه الطريقة أفرزت طريقة بسيطة لتقليص عدد المتغيرات المستقلة وهي خاصية مرغوب فيها خاصة عندما يكون عددها كبيرا وكذلك إذا ما ساعد في زيادة دقة التقدير وهذا ما حدث في المثال المصاحب

---

* كلية العلوم جامعة صنعاء