RT-MAE 9619

STEPWISE PROCEDURES FOR SELECTING VARIABLES IN THE MINIMUM SUM OF ABSOLUTE ERRORS REGRESSION

by

Carmen D.S. André, Silvia N. Elian, Subhash C. Narula
and
Elisete C. Q. Aubin

Palavras-Chave: backward elimination; forward selection; minimum sum of absolute
(Key words) errors regression; stepwise; variables selection.

Classificação AMS: 62J05.
(AMS Classification)

- Setembro de 1996 -

STEPWISE PROCEDURES FOR SELECTING VARIABLES IN THE MINIMUM SUM OF ABSOLUTE ERRORS REGRESSION

Carmen D. S. André
University of Sao Paulo
Sao Paulo, Brazil

Silvia N. Elian
University of Sao Paulo
Sao Paulo, Brazil

Subhash C. Narula
Virginia Commonwealth University
Richmond, Virginia, USA

Elisete C. Q. Aubin
University of Sao Paulo
Sao Paulo, Brazil

Abstract

In the multiple least squares regression, one may use the forward selection, the backward elimination and the stepwise procedures for selecting variables. At present, similar rigorous procedures do not exists for the minimum sum of absolute errors MSAE regression. In this paper, our objective is to propose such procedures for the multiple linear MSAE regression models. We illustrate the proposed procedures with an example.

Key words : backward elimination; forward selection; minimum sum of absolute errors regression; stepwise; variables selection.

1. INTRODUCTION

In several practical problems, multiple linear regression model is often an appropriate model. The initial model, however, may contain a large number of predictor (regressor or independent) variables. It is hoped that this set includes all the relevant variables and their appropriate functions, and at times may include some extraneous variables and their functions. All variables are not equally easy to observe or measure. Moreover, it is possible that only a few of the variables may be enough to explain the data, the phenomenon, or the process under consideration. Clearly, a model with only a few variables is easy to understand and explain. It may also be desirable from computational and statistical considerations. Therefore, it is useful to investigate models with fewer variables. But, we hasten to point out that in most practical problems, as a rule there does not exist a single "best" model but rather many "equally good" models. In selecting the final model, one should use experience, professional judgment in the subject area, and other practical and economic consideration.

For the least squares procedure, several techniques for selecting variables are available, Hocking (1976). The most popular among these algorithms are the stepwise procedures. These procedures select a variable based on its contribution to the regression sum of squares. The tests for including or excluding a variable are well described and documented is several texts, e.g., Draper and Smith (1981), Montgomery and Peck (1992) and Neter, Kutner, Nachtsheim, and Wasserman (1996). However, when the assumptions for the least squares procedure do not hold or the data contain some outliers, the analyst may want to use some other criterion to select a model and estimate its parameters.

The minimum sum of absolute errors MSAE criterion is appropriate for estimating the parameters of the multiple linear regression model whenever the errors follow a long tailed error distribution, or the data contain a few outliers, or the loss function is proportional to the absolute value of the error, Narula and Wellington (1977). The MSAE regression is more robust to outliers in the values of the response variable than the least squares regression, Narula and Wellington (1985). Roodman (1974) and Narula and Wellington (1979) have proposed implicit enumeration algorithms for computing models with one to k variables, where k denotes the total number of variables under consideration. For a computer program to find all possible models using the MSAE criterion, the interested reader may refer to Wellington and Narula (1981).

Stepwise procedures identify a small number of good models with fewer variables by adding or deleting predictor variables. Based on the tests for linear hypotheses for the general linear model by McKean and Schrader (1987), it is possible to develop stepwise procedures for selecting variables in the multiple linear MSAE regression. In this paper, our objective is to develop such procedures. The rest of the paper is organized as follows: In Section 2, we give some preliminary results and develop the stepwise procedures in Section 3. We illustrate the proposed procedures with an example in Section 4 and conclude the paper with a few remarks in Section 5.

2. PRELIMINARY RESULTS

To develop stepwise procedures for variable selection, we need a statistical test, similar to the *partial*-F test used in the least squares regression, to test the contribution of a variable to the regression sum of absolute errors. To do so, we use the test statistic developed by McKean and Schrader (1987) for testing hypotheses in the general linear model and proceed as follows:

Let y denote an n x 1 vector of the values of the response variable corresponding to **X** , an n x (k + 1) matrix of the values of the predictor variables that contains a column of ones for the intercept term. Further, let β, a (k + 1) x 1 vector, denote the unknown parameters and ε , an n x 1 vector, denote the unobservable random errors. Then the multiple linear regression model is

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon. \qquad (1)$$

Let the components of the error vector be mutually independent and identically distributed random variables with a density function f(x) and the scale parameter τ,

$$\tau = (2f(v))^{-1},$$

where v denotes the median of the error distribution.

To test the contribution of variable x_i, given that variables $x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p$ are already in the model, we compute the sum of absolute errors for the models:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon \qquad (2)$$

and

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{i-1} x_{i-1} + \beta_{i+1} x_{i+1} + \ldots + \beta_p x_p + \varepsilon . \qquad (3)$$

Let SAE denote the sum of absolute errors for model (2) with p variables and SAE(i) denote the sum of absolute errors for model (3) without variable x_i. The reduction in the sum of absolute errors is given by SAE(i) - SAE. The test statistics to measure the contribution of variable x_i is

$$F = (\text{SAE(i)} - \text{SAE}) / (\hat{\tau}/2), \qquad (4)$$

where $\hat{\tau}$ is a consistent estimator of τ. The test statistic F has an approximate chi-square distribution with one degree of freedom whenever $H_0 : \beta_i = 0$ is true.

To compute F, one may use any consistent estimator of τ. Let $e_1, e_2, \ldots, e_n$, denote the residuals from the MSAE fit of model (2). Recall that for the MSAE fit, at least (p+1) residuals will be zero. Let $e^*_1, e^*_2, \ldots, e^*_{n^*}$, denote the nonzero residuals where $n^* = n - r$ and r is the number of residuals that are equal to zero. McKean and Schrader (1984) recommend the estimator of τ that is based on the standardized length of a distribution free confidence interval. It is of the form $(e^*_{(m)}, e^*_{(n^*-m+1)})$ where the asymptotic value of m is $((n^* + 1)/2) - z_{\alpha/2}(n^*/4)^{1/2}$ and z_α is the (1 - α) percentile of the standard normal distribution. The estimator $\hat{\tau}$ is:

$$\hat{\tau}_{1-\alpha} = \sqrt{n^*}(e^*_{(n^*-m+1)} - e^*_{(m)}) / (2z_{\alpha/2}) . \qquad (5)$$

They also recommend α = 0.05.

To measure the goodness of fit of a model, McKean and Sievers (1987) proposed two criteria similar to R^2, the coefficient of determination for the least square regression, for the MSAE regression. Let RSAE denote the reduction in sum of absolute errors because of fitting a p variable model, i.e.,

$$\text{RSAE} = \sum_{i=1}^{n} |y_i - median(y_i)| - \text{SAE}, \qquad (6)$$

where SAE is the sum of absolute errors associated with a p variable model. They recommended

$$R_2 = RSAE/(RSAE + (n - p - 1)(\hat{\tau} /2)). \qquad (7)$$

because it is robust.

3. PROPOSED PROCEDURES

Stepwise procedures have been classified in three broad categories: forward selection, backward elimination and stepwise which combines the forward selection and the backward elimination procedures, Montgomery and Peck (1992).

3.1 Forward Selection

The procedure starts with the model $\hat{y}$ = median (y_i) and brings in the variable which corresponds to the maximum value of F in (4) and its contribution is statistically significant at the α level of significance. The next variable to enter the model has the largest value of the test statistic and it is statistically significant at the α level of significance. The process is repeated until no more variable may enter the model. The procedure may be stated as follows:

Step 1F: Compute the sum of absolute errors for the initial model $\hat{y}$ = median (y_i). Go to Step 2F.

Step 2F: Compute the test statistic F in (4) for each variable. Select the variable corresponding to the largest F value and test its significance at the α level of significance. If it is not significant, stop; otherwise, include the variable in the model and go to Step 3F.

Step 3F: Compute the test statistic F in (4) for each of the variable not yet in the model. Select the variable with the largest F value and test its significance at the α level f significance. If it is not significant, stop; otherwise, include the variable in the model and go to Step 4F.

Step 4F: Repeat Step 3F until no new variable may enter the model.

3.2 Backward Elimination

The procedure starts with all the variables in the model. It may be noted that unlike the least squares regression, the multicollinearity among the variables does not cause any computational problems. The F statistic for each variable in the model is computed as if it were the last variable to enter the model. The smallest F value is tested for statistical significance at the α level of significance. If it is not significant, the variable is deleted. Starting with this model, the process is repeated until no more variables can be deleted. The procedure may be stated as follows:

Step 1B: Fit the model with all the variables and compute its sum of absolute errors. Go to Step 2B.

Step 2B: Compute the F value in (4) for each variable as if it were the last variable to enter the model. Select the variable corresponding to the smallest F value and test its significance at the α level of significance. If it is significant, stop; otherwise delete the variable and go to Step 3B.

Step 3B: Starting with the model in Step 2B, compute the F value in (4) for each variable as if it were the last variable to enter the model. Select the variable with to the smallest F value and test its significance at the α level of significance. If it is significant, stop; otherwise delete the variable and go to Step 4B.

Step 4B: Repeat Step 3B until no further variables may be deleted.

3.3 Stepwise Method

This procedure is a combination of the forward selection and the backward elimination procedures. The process starts as with the forward selection process. In fact, the first three steps of the procedure are the same. At the fourth step, the contribution of each variable already in the model is tested in the presence of the last variable to enter the model. This is done by computing F value in (4) for each variable as if it were the last variable to enter and testing it at the α level of significance; if the F value is insignificant, the variable is deleted. This can happen because of the relationship among variables which may make the contribution of a variable insignificant in the presence of other variables. This process is repeated until no more variables can be added to or deleted from the model. In this procedure one may use two levels of significance: one to include a variable and one to delete a variable, i.e., α and β. Some analysts prefer to choose $\alpha = \beta$, although it is not necessary. The procedure may be stated as follows:

Step 1S. Backward Step: Suppose that there are q (≥ 3) variables in the model and let variable xq be the last variable to enter the model. Compute F value in (4) for each variable, x1, ..., xq-1, as if it were the last variable to enter the model. Select the variable corresponding to the smallest F value and test its significance at the α level of significance. If it is significant, go to Step 2S. If it is not significant, delete the variable and repeat this step with the reduced model until a variable can not be deleted, and go to Step 2S.

Step 2S. Forward Step: Starting with the model in Step 1S, compute F value in (4) for each variable not in the model. Select the variable corresponding to the largest F value and test its significance at the α level of significance. If it is not significant, stop; otherwise, include the variable in the model and go to Step 1S.

4. AN EXAMPLE

Interstitial Lung Disease(ILD) refers to a diffuse inflammatory process that occurs predominantly within the interstitial spaces and supporting structures of a lung. Clinical chart and x-rays of a patient with ILD usually suggest an open-chest lung biopsy to establish the diagnosis and to provide additional information on activity and stage of disease. The list of variables and the data from twenty four patients on fourteen variables are given in the Appendix.

Using the forward selection procedure and $\alpha = 0.05$, the selected model is:

$$\hat{y} = 72.22- 7.68\ x_1 + 0.23\ x_2 - 0.06\ x_4 + 0.004\ x_5 - 0.0003\ x6 - 0.08\ x_7 + 124.06\ x_8 -10.67x_{13}\ ,$$

with R_2=0.9166.

The backward elimination procedure with $\alpha = 0.05$ selects the following model:

$$\hat{y} = -24.58 + 4.83\ x_1 + 0.02\ x_4 + 0.040\ x_5 + 0.0018\ x_6 - 0.48\ x_7 + 152.10\ x_8 + 104.05\ x_9 - 13.88\ x_{11} - 7.45\ x_{12} + 6.63\ x_{14},$$

with R_2=0.9540.

The stepwise procedure with $\alpha = 0.05$ for entering a variable to and for deleting a variable from a model results in the following model:

$$\hat{y} = 81.59 - 8.75x_1 + 0.22\ x_2 - 0.06\ x_4 + 95.35\ x_8 - 1.40\ x_{10} - 9.87\ x_{13},$$

with R_2=0.9406.

5. CONCLUDING REMARKS

We have proposed a forward selection, a backward elimination and a stepwise procedure to select variables in the multiple linear MSAE regression model. The procedures are based on a statistical test of hypothesis for the contribution each variable makes to reducing the sum of absolute errors of the model. The statistic R_2 provides one measure of the goodness of fit of a model. As in the least squares regression, the procedures may lead to different models. However, unlike least squares regression, the multicollinearity does not cause any computational problems in the backward elimination procedure. The proposed procedures can be easily implemented on SAS.

ACKNOWLEDGMENT

We thank Dr. Paulo H. N. Saldiva, Faculdade de Medicina da Universidade de Sao Paulo, for the data for the example. We also thank FAPESP for the grant that partially supported the work of Professor Narula.

REFERENCES

Draper, N. R. and Smith, H. (1981). *Applied Linear Regression, Second Edition.* John Wiley and Sons, New York, N.Y.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics, 32*, 1-49.

McKean, J. W. and Schrader, R. M. (1984). A comparison of methods for studentizing the sample median. *Communications in Statistics - Simulation and Computations, B13*, 751-773.

McKean, J. W. and Schrader, R. M. (1987). Least absolute errors analysis of variance. *Statistical Data Analysis Based on the L_1-Norm and Related Methods* (Y. Dodge, editor), Elsevier Science Publishers B.V., 297 - 305.

McKean, J. W. and Sievers, G. L. (1987). Coefficient of determination for least absolute deviation analysis. *Statistics and Probability Letters, 5*, 49-54.

Montgomery, D. C. and Peck, E. A. (1982). *Introduction to Linear Regression Analysis, Second Edition.* John Wiley and Sons, New York, N. Y.

Narula, S. C. and Wellington, J. W. (1977). Prediction, linear regression and minimum sum of absolute errors. *Technometrics, 19*, 185-190.

Narula, S. C. and Wellington, J. W. (1979). Selection of variables in linear regression using the minimum sum of weighted absolute errors criterion. *Technometrics, 21*, 299-306.

Narula, S. C. and Wellington, J. W. (1985). Interior analysis for the minimum sum of absolute errors regression. *Technometrics, 27*, 181-188.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996). *Applied Linear Statistical Models.* Richard D. Irwin, Inc., Chicago.

Roodman, G. M. (1974) A procedure for optimal stepwise MSAE regression analysis. *Operations Research, 22*, 393-399.

Wellington, J. W. and Narula, S. C. (1981). Variable selection in multiple linear regression using the minimum sum of weighted absolute errors criterion. *Communications in Statistics - Simulation and Computations, B10*, 641-648.

APPENDIX

THE LIST OF VARIABLES

Response variable:

y: forced vital capacity

Predictor variables:

x_1: sex: 0 = Male and 1 = Female;
x_2: age (in years);
x_3: smoking: 0 = Smoker and 1 = Nonsmoker;
x_4: Area fraction of epitelial cells/10 000 μm^2 of alveolar tissue;
x_5: Area fraction of fusiform cells/10 000 μm^2 of alveolar tissue;
x_6: Area fraction of mononucleated cells/10 000 μm^2 of alveolar tissue;
x_7: Area fraction of polymorphonuclear cells /10 000 μm^2 of alveolar tissue;
x_8: Total cellularity/10 000 μm^2 of alveolar tissue;
x_9: Area fraction of capillaries/10 000 μm^2 of alveolar tissue;
x_{10}: Score of bronchiolitis obliterans (zero to four);
x_{11}: Score of smooth muscle (zero to four);
x_{12}: Score of vascular sclerosis (zero to four);
x_{13}: Score of honeycombing (zero to four);
x_{14}: Score of intra alveolar cell isquamation (zero to four);

THE DATA SET

obs	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}
1.	56	1	64	0	192.405	359.71	669.24	0.000	0.231	0.289	3	4	1	4	2
2.	75	2	39	0	398.588	441.53	163.06	20.706	0.251	0.578	0	0	3	0	0
3.	32	2	39	0	671.674	622.29	1728.57	49.308	0.043	0.203	3	0	0	0	0
4.	88	1	69	1	227.424	539.19	145.42	13.424	0.153	0.615	0	0	0	0	0
5.	83	1	41	0	310.136	419.39	88.11	3.525	0.143	0.551	0	0	0	0	0
6.	59	1	42	1	187.597	378.95	82.54	1.251	0.150	0.785	0	4	3	3	2
7.	51	1	32	1	405.836	411.85	261.54	30.062	0.225	0.240	0	0	0	0	0
8.	67	1	45	1	100.237	346.53	223.38	2.864	0.183	0.725	0	0	3	1	2
9.	60	2	53	0	144.290	397.77	129.99	0.000	0.176	0.696	3	1	4	2	0
10.	98	1	46	1	149.187	275.22	204.49	14.147	0.251	0.577	0	0	2	0	4
11.	48	2	44	0	211.614	398.81	278.35	4.883	0.174	0.703	0	1	3	0	2
12.	82	1	44	0	254.398	376.39	297.54	7.439	0.242	0.593	3	2	0	0	2
13.	86	2	57	0	167.728	384.79	4624.07	3.289	0.203	0.702	0	1	2	0	0
14.	103	2	49	0	337.145	597.76	614.08	12.410	0.313	0.554	0	0	2	0	0
15.	115	2	65	0	276.864	365.31	401.66	8.206	0.206	0.572	0	0	2	0	3
16.	64	2	26	0	309.206	512.22	99.65	27.510	0.224	0.579	0	0	3	0	1
17.	57	1	46	1	173.373	367.14	308.02	24.222	0.204	0.722	0	2	3	3	2
18.	82	1	28	1	238.277	375.29	223.85	64.037	0.178	0.685	0	0	2	0	1
19.	50	2	52	1	130.308	374.79	423.90	34.747	0.175	0.697	3	2	3	3	2
20.	48	1	49	1	165.546	318.45	284.34	37.911	0.203	0.674	4	2	2	4	2
21.	57	2	32	0	168.547	394.52	282.60	1.349	0.165	0.647	0	2	4	2	2
22.	45	1	57	0	621.861	1477.29	416.57	151.746	0.238	0.685	2	3	2	2	2
23.	77	1	72	0	607.268	171.91	2529.51	89.094	0.468	0.435	0	0	2	2	2
24.	92	1	57	1	404.735	1443.59	2022.71	93.677	0.293	0.618	0	0	3	0	0

ÚLTIMOS RELATÓRIOS TÉCNICOS PUBLICADOS

9601 - MENTZ, R.P.; MORETTIN, P.A. and TOLOI, C.M.C. Bias correction for estimators of the residual variance in the ARMA (1,1) Model. São Paulo, IME-USP, 1996. 21p. (RT-MAE-9601)

9602 - SINGER, J.M.; PERES, C.A.; HARLE, C.E. Performance of Wald's test for the Hardy-Weiberg equilibrium with fixed sample sizes. São Paulo, IME-USP, 1996.15p. (RT-MAE-9602).

9603 - SINGER, J.M. and E. SUYAMA, E. Dispersion structure, Hierarchical models, Random effects models, Repeated measures. São Paulo, IME-USP, 1996. 21p. (RT-MAE-9603).

9604 - LIMA, A.C.P. and SEN, P.K. A Matrix-Valued Counting Process with First-Order Interactive Intensities. São Paulo, IME-USP, 1996, 25p. (RT-MAE-9604)

9605 - BOTTER, D.A. and SINGER, J.M. Experimentos com Intercâmbio de Dois Tratamentos e Dois Períodos: Estratégias para Análise e Aspectos Computacionais, São Paulo, IME-USP, 1996. 18p. (RT-MAE-9605)

9606- MORETTIN, P.A. From Fourier to Wavelet Analysis of Time Series. São Paulo, IME-USP, 1996. 11p. (RT-MAE-9606)

9607 - SINGER, J.M.; HO, L.L. Regression Models for Bivariate Counts. São Paulo, IME-USP, 1996. 20p. (RT-MAE-9607)

9608 - CORDEIRO, G.M. and FERRARI, S.L.P. A method of mements for finding Bartlett-type corrections. São Paulo, IME-USP, 1996. 11p. (RT-MAE-9608)

9609 - VILCA-LABRA, F.; ARELLANO-VALLE, R.B.; BOLFARINE, H. Elliptical functional models. São Paulo, IME-USP, 1996. 20p. (RT-MAE-9609)

9610 - BELITSKY, V.; FERRARI, P.A.; KONNO, N. A Refinement of Harris-FKG Inequality for Oriented Percolation. São Paulo, IME-USP, 1996. 13p. (RT-MAE-9610)

9611 - GIMENEZ, P.; BOLFARINE, H. Unbiased Score Functions in Error-In-Variables Models. São Paulo, IME-USP, 1996. 23p. (RT-MAE-9611)

9612 - BUENO, V.C. Comparing component redundancy allocation in K-out-of-n system. IME-USP, 1996. 10p. (RT-MAE-9612)

9613 - GALEA, M.; PAULA, G.A.; BOLFARINE, H. Local influence in elliptical linear regression models. IME-USP, 1996. 14p. (RT-MAE-9613)

9614 - LIMA, C.R.; BOLFARINE, H.; SANDOVAL, M.C. Linear calibration in multiplicative measurement error models. IME-USP, 1996. 12p. (RT-MAE-9614)

9615 - AVERBACH, M.; CUTAIT, R.; WECHSLER, S.; CORRÊA, P.; BORGES, J.L.A. Aplicação da inferência bayesiana no estudo das probabilidades de diagnóstico, por colonoscopia, das afecções colorretais em portadores da síndrome da imunodeficiência adquirida com diarréia. IME-USP, 1996. 10p. (RT-MAE-9615)

9616 - YOSHIDA, O.S.; LEITE, J.G.; BOLFARINE, H. Inferência bayesiana do número de espécies de uma população. IME-USP, 1996. 33p. (RT-MAE-9616)

9617 - CARMONA, S.C.; TANAKA, N.I. Exponential estimates for "Not Very Large Deviations" and ware front propagation for a class of reaction-diffusion equations. IME-USP, 1996. 30p. (RT-MAE-9617)

9618 - IRONY, T.Z.; PEREIRA, C.A.B.; TIWARI, R.C. On The Comparison Between Two Correlated Proportions in 2 x 2 Tables. IME-USP, 1996. 15p. (RT-MAE-9618)

The complete list of "Relatórios do Departamento de Estatística", IME-USP, will be sent upon request.

Departamento de Estatística
IME-USP
Caixa Postal 66.281
05389-970 - São Paulo, Brasil