Recognition of Subsets of Informative Variables in Regression

N.A. Nechval [1], K.N. Nechval [2], M. Purgailis [1], U. Rozevskis [1], V. Strelchonok [3], M. Moldovan [4], I. Bausova [1], D. Skiltere [1]

1) University of Latvia, Raina Blvd 19, LV-1050 Riga, Latvia, nechval@junik.lv

2) Transport Institute, Lomonosov Street 1, LV-1019 Riga, Latvia, konstan@tsi.lv

3) Baltic International Academy, Lomonosov Street 1, LV-1019 Riga, Latvia, str@apollo.lv

4) University of Melbourne, VIC-3010 Melbourne, Australia, max.moldovan@gmail.com

Abstract: *A new approach is proposed to address the subset recognition problem in multiple linear regression, where the objective is to recognize a minimal subset of predictor variables without sacrificing any explanatory power. A parameter stability solution of this approach yields a number of informative subsets. To obtain this solution, new parameter stability criteria are repeatedly used. The subsets generated are compared to ones generated by several standard procedures. The results suggest that the new approach finds subsets that compare favorably against the standard procedures in terms of the generally accepted measure:* R^2.

***Keywords*:** Regression, variables, subset recognition.

1. INTRODUCTION

A number of studies in the statistical literature discuss the problem of selecting (recognizing) the best subset of predictor variables in regression. Such studies focus on subset selection methodologies, selection criteria, or a combination of both. The traditional selection methodologies can be enumerative (e.g. all subsets and best subsets procedures), sequential (e.g. forward selection, backward elimination, stepwise regression, and stagewise regression procedures), and screening-based (e.g. ridge regression and principal components analysis). Standard texts like Draper and Smith [1] and Montgomery and Peck [2] provide clear descriptions of these methodologies.

Some of the reasons for using only a subset of the available predictor variables (given by Miller [3]) are

- to estimate or predict at a lower cost by reducing the number of variables on which data are to be collected;
- to predict more accurately by eliminating uninformative variables;
- to describe multivariate data sets parsimoniously; and
- to estimate regression coefficients with smaller standard errors (particularly when some of the predictors are highly correlated).

These objectives are of course not completely compatible. Prediction is probably the most common objective, and here the range of values of the predictor variables for which predictions will be required is important. The subset of variables giving the best predictions in some sense, averaged over the region covered by the calibration data, may be very inferior to other subsets for extrapolation beyond this region. For prediction purposes, the regression coefficients are not the primary objective, and poorly estimated coefficients can sometimes yield acceptable predictions. On the other hand, if process control is the objective then it is of vital importance to know accurately how much change can be expected when one of the predictors changes or is changed.

Suppose that $\mathbf{Y}$, a variable of interest, and $\mathbf{X}_1, ..., \mathbf{X}_v$, a set of potential explanatory variables or predictors, are vectors of n observations. The problem of variable selection, or subset selection (recognition) as it is often called, arises when one wants to model the relationship between $\mathbf{Y}$ and a subset of $\mathbf{X}_1, ..., \mathbf{X}_v$, but there is uncertainty about which subset to use. Such a situation is particularly of interest when v is large and $\mathbf{X}_1, ..., \mathbf{X}_v$ is thought to contain many redundant or irrelevant variables.

The variable selection problem is most familiar in the linear regression context, where attention is restricted to normal linear models. Letting w index the subsets of $\mathbf{X}_1, ..., \mathbf{X}_v$ and letting p_w be the number of the parameters of the model based on the wth subset, the problem is to select and fit a model of the form

$$\mathbf{Y} = \mathbf{X}_w \boldsymbol{\beta}_w + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{X}_w$ is an $n \times p_w$ matrix whose columns correspond to the wth subset, $\boldsymbol{\beta}_w$ is a $p_w \times 1$ vector of regression coefficients, and $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$. More generally, the variable selection problem is a special case of the model selection problem where each model under consideration corresponds to a distinct subset of $\mathbf{X}_1, ..., \mathbf{X}_v$. Typically, a single model class is simply applied to all possible subsets.

The fundamental developments in variable selection seem to have occurred directly in the context of the linear model (1). Historically, the focus began with the linear model in the 1960s, when the first wave of important developments occurred and computing was expensive. The focus on the linear model still continues, in part because its analytic tractability greatly facilitates insight, but also because many problems of interest can be posed as linear variable selection problems. For example, for the problem of non-parametric function estimation, $\mathbf{Y}$ represents the values of the unknown function, and $\mathbf{X}_1, ..., \mathbf{X}_v$ represent a linear basis, such as a wavelet basis or a spline basis.

One of the fascinating aspects of the variable selection problem has been the wide variety of methods that have been brought to bear on the problem. Because of space limitations, it is of course impossible to even mention them all, and so we focus on only a few to illustrate the general thrust of developments. An excellent and comprehensive treatment of variable selection methods prior to 1990 was provided by Miller [3]. As we discuss, many promising new approaches have appeared over the last decade.

A distinguishing feature of variable selection problems is their enormous size. Even with moderate values of v, computing characteristics for all 2^v models is

prohibitively expensive, and some reduction of the model space is needed. Focusing on the linear model (1), early suggestions based such reductions on the residual sum of squares, which provided a partial ordering of the models. Taking advantage of the chain structure of subsets, branch and bound methods such as the algorithm of Furnival and Wilson [4] were proposed to logically eliminate large numbers of models from consideration. When feasible, attention was often restricted to the "best subsets" of each size. Otherwise, reduction was obtained with variants of stepwise methods that sequentially add or delete variables based on greedy considerations (e.g., Efroymson [5]). Even with advances in computing technology, these methods continue to be the standard workhorses for reduction.

Once attention was reduced to a manageable set of models, criteria were needed for selecting a subset model. The earliest developments of such selection criteria, again in the linear model context, were based on attempts to minimize the mean squared error of prediction. Different criteria corresponded to different assumptions about which predictor values to use, and whether they were fixed or random (see Hocking [6]; Thompson [7] and the references therein). Perhaps the most familiar of those criteria is the Mallows

$$C_p = \frac{\mathrm{RSS}_w}{\hat{\sigma}^2_{\mathrm{full}}} + 2p_w - n, \qquad (2)$$

where RSS_w is the residual sum of squares for the model based on the wth subset and $\hat{\sigma}^2_{\mathrm{full}}$ is the usual unbiased estimate of σ^2 based on the full model. The standard texts, such as Draper and Smith [1], Montgomery and Peck [2] and Myers [8], recommend plotting C_p, against p for all possible regressions and choosing an equation with low C_p or with C_p close to p. If σ^2 is known, any model which provides unbiased estimates of the regression coefficients, i.e. which contains all important regressors, has $E(C_p) = p$. Two of the other most popular criteria, motivated from very different viewpoints, are the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Letting $\hat{L}_w$ denote the maximum log-likelihood of the wth model, AIC selects the model that maximizes $(\hat{L}_w - p_w)$, whereas BIC selects the model that maximizes $(\hat{L}_w - (\log n)p_w/2)$. Akaike [9] motivated AIC from an information theoretic standpoint as the minimization of the Kullback-Leibler distance between the distributions of Y under the wth model and under the true model. To lend further support, an asymptotic equivalence of AIC and cross-validation was shown by Stone [10]. In contrast, Schwarz [11] motivated BIC from a Bayesian standpoint, by showing that it was asymptotically equivalent (as $n \to \infty$) to selection based on Bayes factors. BIC was further justified from a coding theory viewpoint by Rissanen [12].

Comparisons of the relative merits of AIC and BIC based on asymptotic consistency (as $n \to \infty$) have flourished in the literature. As it turns out, BIC is consistent when the true model is fixed (Haughton [13]), whereas AIC is consistent if the dimensionality of the true model increases with n (at an appropriate rate) (Shibata [14]). Stone [15] provided an illuminating discussion of these two viewpoints.

For the linear model (1), many of the popular selection criteria are special cases of a penalized sum of squares criterion, providing a unified framework for comparisons. Assuming σ^2 known to avoid complications, this general criterion selects the subset model that minimizes

$$\frac{\mathrm{RSS}_w}{\sigma^2} + cp_w, \qquad (3)$$

where c is a preset "parametric dimensionality penalty." Intuitively, (3) penalizes RSS_w/σ^2 by c times p_w, the parametric dimension of the wth model. AIC and minimum C_p are essentially equivalent, corresponding to $c = 2$, and BIC is obtained by setting $c = \log n$. By imposing a smaller penalty, AIC and minimum C_p will select larger models than BIC (unless n is very small).

Further insight into the choice of c is obtained when all of the predictors are orthogonal, in which case (3) simply selects all of those predictors with T-statistics t for which $t^2 > c$. When $\mathbf{X}_1, \ldots, \mathbf{X}_v$ are in fact all unrelated to $\mathbf{Y}$ (i.e., the full model regression coefficients are all 0), AIC and minimum C_p are clearly too liberal and tend to include a large proportion of irrelevant variables. A natural conservative choice for c, namely $c = 2\log v$, is suggested by the fact that under this null model, the expected value of the largest squared T-statistic is approximately $2\log v$ when v is large. This choice is the risk inflation criterion (RIC) proposed by Foster and George [16] and the universal threshold for wavelets proposed by Donoho and Johnstone [17]. Both of these articles motivate $c = 2\log v$ as yielding the smallest possible maximum inflation in predictive risk due to selection (as $v \to \infty$), a minimax decision theory standpoint. Motivated by similar considerations, Tibshirani and Knight [18] recently proposed the covariance inflation criterion (CIC), a nonparametric method of selection based on adjusting the bias of in-sample performance estimates. Yet another promising adjustment based on a generalized degrees of freedom concept was proposed by Ye [19].

Many other interesting criteria corresponding to different choices of c in (3) have been proposed in the literature (see, e.g., Hurvitz and Tsai [20-21]; Rao and Wu [22]; Shao [23]; Wei [24]; Zheng and Loh [25] and the references therein). One of the drawbacks of using a fixed choice of c is that models of a particular size are favored; small c favors large models, and large c favors small models. Adaptive choices of c to mitigate this problem have been recommended by Benjamini and Hochberg [26], Clyde and George [27-28], Foster and George [16], Johnstone and Silverman [29].

An alternative to explicit criteria of the form (3), is selection based on predictive error estimates obtained by intensive computing methods such as the bootstrap (e.g., Efron [30]; Gong [31]) and cross-validation (e.g., Shao [32]; Zhang [33]). An interesting variant of these is the little bootstrap (Brieman [34]), which estimates the predictive error of selected models by mimicking replicate data comparison. The little bootstrap compares

favorably to selection based on minimum C_p or the conditional bootstrap, whose performances are seriously denigrated by selection bias.

Another drawback of traditional subset selection methods, which is beginning to receive more attention, is their instability relative to small changes in the data. Two novel alternatives that mitigate some of this instability for linear models are the nonnegative garrotte (Brieman [35]) and the lasso (Tibshirani [36]). Both of these procedures replace the full model least squares criterion by constrained optimization criteria. As the constraint is tightened, estimates are zeroed out, and a subset model is identified and estimated.

The fully Bayesian approach to variable selection is as follows (George [37]). For a given set of models $M(1)$, ..., $M(2^v)$, where $M(w)$ corresponds to the wth subset of $\mathbf{X}_1$, ..., $\mathbf{X}_v$, one puts priors $\pi(\boldsymbol{\beta}(w)|M(w))$ on the parameters of each $M(w)$ and a prior on the set of models $\pi(M(1))$, ..., $\pi(M(2^v))$. Selection is then based on the posterior model probabilities $\pi(M(w)|\mathbf{Y})$, which are obtained in principle by Bayes's theorem.

Although this Bayesian approach appears to provide a comprehensive solution to the variable selection problem, the difficulties of prior specification and posterior computation are formidable when the set of models is large. Even when v is small and subjective considerations are not out of the question (Garthwaite and Dickey [38]), prior specification requires considerable effort.

2. CRITERIA FOR RECOGNITION OF SUBSETS OF INFORMATIVE VARIABLES

Parameter Stability Criterion (PSC). This criterion (denoted by PSC) is given by

$$\mathrm{PSC} = \mathrm{RSS}_{M(w)}, \left| \frac{\hat{a}_i}{s_{\hat{a}_i}} \right| > t_{k;\alpha}, \forall i = 1(1) p_{M(w)}, \tag{4}$$

where $\mathrm{RSS}_{M(w)}$ is the residual sum of squares for the wth subset model $M(w)$, which has the number of parameters equal to $p_{M(w)}$, $\hat{a}_i$ is an estimate of the parameter a_i of the model $M(w)$, $s_{\hat{a}_i}$ represents the estimated standard deviation of $\hat{a}_i$, $\hat{a}_i / s_{\hat{a}_i}$ follows the Student distribution (T-distribution) with $k = n - p_{M(w)}$ degrees of freedom, n is the number of observations, $t_{k;\alpha}$ is an upper-tail value of the T-statistic at the given significance level α, i.e., $\Pr\{T > t_{k;\alpha}\} = \alpha$.

According to (4), the best model (subset of informative variables) denoted by $M^*(w)$ is determined as

$$\begin{aligned} M^*(w) &= \arg \inf_{M(w) \in \{M(w): w \in \{w\}\}} \mathrm{RSS}_{M(w)} \\ &= \arg \inf_{M(w) \in \{M(w): w \in \{w\}\}} \frac{\mathrm{RSS}_{M(w)}}{\mathrm{TSS}} \\ &= \arg \inf_{M(w) \in \{M(w): w \in \{w\}\}} \left(1 - R^2_{M(w)}\right) \\ &= \arg \sup_{M(w) \in \{M(w): w \in \{w\}\}} R^2_{M(w)} \end{aligned} \tag{5}$$

subject to

$$\left| \frac{\hat{a}_i}{s_{\hat{a}_i}} \right| > t_{k;\alpha}, \forall i = 1(1) p_{M(w)}, \tag{6}$$

where the coefficient of determination $R^2_{M(w)}$ ($0 \le R^2_{M(w)} \le 1$) for the wth subset model $M(w)$ is computed as

$$R^2_{M(w)} = 1 - \frac{\mathrm{RSS}_{M(w)}}{\mathrm{TSS}}, \tag{7}$$

TSS is the total sum of squares.

This criterion involves the data fit indicator (5) and parameter stability indicator (6). It allows one to recognize the suitable stable subset model minimizing the residual sum of squares.

Multiplicative Parameter Stability Criterion (MPSC). This criterion (denoted by MPSC) is given by

$$\begin{aligned} &\mathrm{MPSC} \\ &= \ln\left(\frac{1}{1 - R^2_{M(w)}}\right) \frac{1}{\sqrt{p_{M(w)}}}, \left| \frac{\hat{a}_i}{s_{\hat{a}_i}} \right| > t_{k;\alpha}, \forall i = 1(1) p_{M(w)}. \end{aligned} \tag{8}$$

According to (8), the best subset model $M^*(w)$ is determined as

$$M^*(w) = \arg \sup_{M(w) \in \{M(w): w \in \{w\}\}} \left(\ln\left(\frac{1}{1 - R^2_{M(w)}}\right) \frac{1}{\sqrt{p_{M(w)}}} \right) \tag{9}$$

subject to

$$\left| \frac{\hat{a}_i}{s_{\hat{a}_i}} \right| > t_{k;\alpha}, \forall i = 1(1) p_{M(w)}. \tag{10}$$

This criterion involves the parametrically penalized data fit indicator (9) and the parameter stability indicator (10). It allows one to recognize the suitable stable subset model at a lower cost by reducing the number of variables on which data are to be collected.

Modified Multiplicative Parameter Stability Criterion (MMPSC). This criterion (denoted by MMPSC) is given by

$$\begin{aligned} &\mathrm{MMPSC} = \ln\left(\frac{1}{1 - R^2_{M(w)}}\right) \frac{\ln \Sigma_{M(w)}}{\sqrt{p_{M(w)}}}, \\ &\left| \frac{\hat{a}_i}{s_{\hat{a}_i}} \right| > t_{k;\alpha}, \forall i = 1(1) p_{M(w)}, \end{aligned} \tag{11}$$

where

$$\Sigma_{M(w)} = \sum_{i=1}^{p_{M(w)}} \left| \frac{\hat{a}_i}{s_{\hat{a}_i}} \right|. \tag{12}$$

According to (11), the best subset model $M^*(w)$ is determined as

$$M^*(w) = \arg \sup_{M(w) \in \{M(w): w \in \{w\}\}} \left(\ln\left(\frac{1}{1 - R^2_{M(w)}}\right) \frac{\ln \Sigma_{M(w)}}{\sqrt{p_{M(w)}}} \right) \tag{13}$$

subject to

$$\left| \frac{\widehat{a}_i}{s_{\widehat{a}_i}} \right| > t_{k;\alpha}, \forall i = 1(1) p_{M(w)}. \quad (14)$$

This criterion involves the parametrically penalized data fit indicator (9) and parameter stability indicator (10). It allows one to recognize the most stable subset model at a lower cost by reducing the number of variables on which data are to be collected.

Power Parameter Stability Criterion (PPSC). This criterion (denoted by PPSC) is given by

$$\text{PPSC} = \left[1 - R^2_{M(w)}\right]^{\frac{1}{\sqrt{p_{M(w)}}}}, \left| \frac{\widehat{a}_i}{s_{\widehat{a}_i}} \right| > t_{k;\alpha}, \forall i = 1(1) p_{M(w)}, \quad (15)$$

According to (15), the best subset model is determined as

$$M^*(w) = \arg \inf_{M(w) \in \{M(w): w \in \{w\}\}} \left[1 - R^2_{M(w)}\right]^{\frac{1}{\sqrt{p_{M(w)}}}} \quad (16)$$

subject to

$$\left| \frac{\widehat{a}_i}{s_{\widehat{a}_i}} \right| > t_{k;\alpha}, \forall i = 1(1) p_{M(w)}. \quad (17)$$

This criterion involves the parametrically penalized data fit indicator (15) and the parameter stability indicator (17). It allows one to recognize the suitable stable subset model at a lower cost by reducing the number of variables on which data are to be collected.

Modified Power Parameter Stability Criterion (MPPSC). This criterion (denoted by MPPSC) is given by

$$\text{MPPSC} = \left[1 - R^2_{M(w)}\right]^{\frac{\ln \Sigma_{M(w)}}{\sqrt{p_{M(w)}}}}, \left| \frac{\widehat{a}_i}{s_{\widehat{a}_i}} \right| > t_{k;\alpha}, \forall i = 1(1) p_{M(w)}, \quad (18)$$

According to (8), the best subset model is determined as

$$M^* = \arg \inf_{M(w) \in \{M(w): w \in \{w\}\}} \left[1 - R^2_{M(w)}\right]^{\frac{\ln \Sigma_{M(w)}}{\sqrt{p_{M(w)}}}} \quad (19)$$

subject to

$$\left| \frac{\widehat{a}_i}{s_{\widehat{a}_i}} \right| > t_{k;\alpha}, \forall i = 1(1) p_{M(w)}. \quad (20)$$

This criterion involves the penalized data fit indicator (18) and the parameter stability indicator (20). It allows one to recognize the most stable subset model at a lower cost by reducing the number of variables on which data are to be collected.

3. EXAMPLES

Example 1: *Hald cement data*. Montgomery and Peck [2] (pp. 256-266) illustrated variable selection techniques on the Hald cement data and gave several references to other analyses. The data are shown in Table 1.

Table 1. The Hald cement data.

i	y_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	6
8	72.5	1	31	22	44
9	93.1	2	54	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34
12	113.3	11	66	9	12
13	109.4	10	68	8	12

The used data code is as follows: x_1 = amount of tricalcium aluminate, x_2 = amount of tricalcium cilicate, x_3 = amount of tetracalcium alumino ferrite, x_4 = amount of dicalcium cilicate; y = heat evolved in calories per gram of cement. The response variable is the heat evolved y in a cement mix, and the four explanatory variables are ingredients in the mix. When a linear model

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_4 + \varepsilon \quad (21)$$

is fitted, the residuals show no evidence of any problems. But an important feature of these data is that the variables x_1 and x_3 are highly correlated (r_{13} = - 0.824), as are the variables x_2 and x_4 (with r_{24} = - 0.973). Thus we would expect any subset w of $\{x_1, x_2, x_3, x_4\}$ that includes one variable from a highly correlated pair would do as well as any subset that also includes the other member.

MPSC and PPSC select (at the given significance level α = 0.05) the subset model $M^*(w)$, which is given by

$$y = a_0 + a_1 x_1 + a_2 x_2 + \varepsilon, \quad (22)$$

where $w = \{x_1, x_2\}$ and $p_{M(w)} = 3$. It will be noted that the algorithm of Efroymson [5] gives the very same result but via more complex way. MMPSC and MPPSC select (at the given significance level α = 0.05) the subset model $M^*(w)$, where $w = \{x_1, x_4\}$ and $p_{M(w)} = 3$. It should be remarked that the more complex algorithm proposed in [39] gives the very same result. PSC final choice is $M^*(w)$ with $w = \{x_1, x_3, x_4\}$ and $p_{M(w)} = 4$.

Example 2: *Hudson data*. The data set (x_i, y_i), i=1(1)19, analyzed here was simulated using the model:

$$y_i = 1 + x_i - 0.55 x_i^2 + 0.001 x_i^3 + \varepsilon_i, \quad (23)$$

where ε_i, i=1(1)19, are independent and normal with mean zero and variance 1. The data taken from [40] are presented in Table 2.

Table 2. The Hudson data.

i	x_i	y_i	i	x_i	y_i
1	2	2.84	11	22	7.35
2	4	5.50	12	24	6.11
3	6	5.96	13	26	6.67
4	8	4.50	14	28	9.67
5	10	6.45	15	30	7.35
6	12	7.39	16	32	9.99
7	14	6.67	17	34	10.31
8	16	5.72	18	36	12.03
9	18	7.95	19	38	13.51
10	20	5.93			

Assuming that a model of the data belongs to the class of models,

$$y = a_0 + a_1 x + a_2 x^2 + \ldots + a_k x^k + \varepsilon, \quad k \geq 1, \qquad (24)$$

the final choice of PSC, MPSC, MMPSC, PPSC and MPPSC of the best model is k=3, true degree, i.e.,

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \varepsilon. \qquad (25)$$

It will be noted that Hudson obtained the very same result using more complex technique. The Hudson data with the best regression curve are shown in Fig. 1.

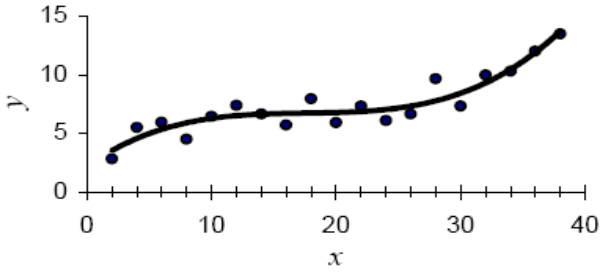

Fig. 1 – The Hudson data with the best regression curve.

Example 3: *Steam data* (Draper and Smith [1], App. A). The used data code is as follows: x_1 = pounds of real fatty acid in storage per month, x_2 = pounds of crude glycerine made, x_3 = average wind velocity in miles per hour, x_4 = calendar days per month, x_5 = operating days per month, x_6 = days below 32°F, x_7 = average atmospheric temperature, degrees F, x_8 = average wind velocity, x_9 = number of startups; y = pounds of steam used monthly.

PSC, MPSC, MMPSC, PPSC and MPPSC select (at the given significance level α = 0.05) the subset model $M^*(w)$ with $w = \{x_2, x_3, x_7\}$ and $p_{M(w)} = 4$, which fits better than $M(w)$ with $w = \{x_1, , x_7\}$ and $p_{M(w)} = 3$ found using the more complex algorithm proposed in [39].

Example 4: *Simulated data*. The data set (x_i, y_i), i = 1(1)100 analyzed here was simulated using the model:

$$y_i = 0.3 + 2x_i - 5x_i^2 + 3x_i^3 + \varepsilon_i, \qquad (26)$$

where, for i=1(1)100, x_i=i/100 and ε_i are independent and normal with mean zero and variance 0.15^2. The situation is such that the true model is known to belong to the class of models given by (24).

The simulation data are shown, with the true regression curve, in Fig. 2.

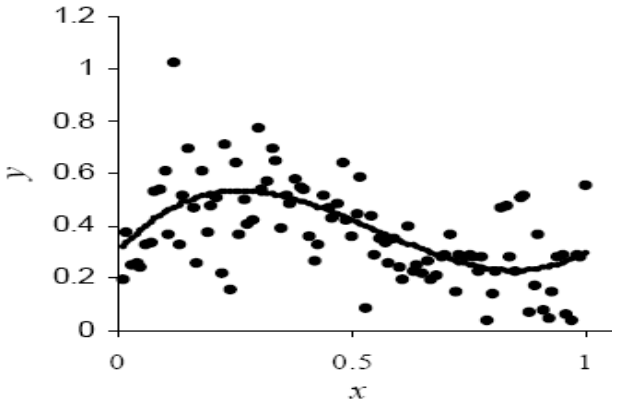

Fig. 2 – Simulated data set with the true regression curve.

BIC, PSC, MPSC, MMPSC, PPSC and MPPSC choose k=3, the true degree. AIC's final choice is k= 8, a clear overfitting.

4. CONCLUSIONS

Subset selection in multiple linear regression is a problem of great practical importance. There are various methods for subset selection and various selection criteria. While there is no clear consensus regarding which method is the best and which criterion is the most appropriate, there is a general agreement an effective method is needed.

Clearly, this paper does not put to rest the question about which is the best subset selection method. However, the proposed approach has certain advantages. First, it quickly produces a reasonable number of subsets having the desirable quality. Compared to the standard sequential procedures that come up with a single "best" model, the proposed approach provides the analyst with a set of "best" models lying on the efficient frontier. The analyst has the option of comparing these solutions with respect to his or her own experience in the specific context and also with respect to other statistical criteria. Thus, the proposed approach gives the analyst the flexibility to pick the best among the best.

Today, variable selection procedures are an integral part of virtually all widely used statistics packages, and their use will only increase as the information revolution brings us larger datasets with more and more variables. The demand for variable selection will be strong, and it will continue to be a basic strategy for data analysis.

5. ACKNOWLEDGMENTS

This research was supported in part by Grant No. 06.1936 and Grant No. 07.2036 from the Latvian Council of Science and the National Institute of Mathematics and Informatics of Latvia.

6. REFERENCES

[1] N.R. Draper, H. Smith. *Applied Regression Analysis*, 2nd edn. Wiley, New York, 1981.

[2] D.C. Montgomery, E.A. Peck. *Introduction to Linear Regression Analysis*, 2nd edn. Wiley, New York, 1992.

[**3**] A. Miller. *Subset Selection in Regression*. Chapman and Hall, London, 1990.

[4] G.M. Furnival, R.W. Wilson. Regression by leaps and bounds, *Technometrics* 16 (1974), pp. 499-511.

[5] M.A. Efroymson. Multiple regression analysis. In: Ralston, A., Wilf, H. S. (eds.) *Mathematical Methods for Digital Computers*. Wiley, New York, 1960, pp. 191-203.

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

[7] M.L. Thompson. Selection of variables in multiple regression: Part I. A review and evaluation, *International Statistical Review* 46 (1978), pp. 1-19.

[8] R.L. Myers. *Classical and Modem Regression Analysis*, 2nd edn. Wiley, New York, 1992.

[9] H. Akaike. Information theory and an extension of the maximum likelihood principle. In: Petrov, B. N., Csaki, F. (eds.) *Proc. of the 2nd International Symposium on Information Theory*, 1973. Akademia Kiado, Budapest, 1973, pp. 267-281.

[10] M. Stone. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, *Journal of the Royal Statistical Society*, Ser. B 39 (1977), pp. 44-47.

[11] G. Schwarz. Estimating the dimension of a model, *The Annals of Statistics* 6 (1978), pp. 461-464.

[12] J. Rissanen. Modeling by shortest data description, *Automatica* 14 (1978), pp. 465-471.

[13] D. Haughton. On the choice of a model to fit data from an exponential family, *The Annals of Statistics* 16 (1988), pp. 342-355.

[14] R. Shibata. An optimal selection of regression variables, *Biometrika* 68 (1981), pp. 45—54.

[15] M. Stone. Comments on model selection criteria of Akaike and Schwarz, *Journal of the Royal Statistical Society*, Ser. B 41 (1979), pp. 276—278.

[16] D.P. Foster, E.I. George. The risk inflation criterion for multiple regression, *The Annals of Statistics* 22 (1994), pp. 1947-1975.

[17] D.L. Donoho, I.M. Johnstone. Ideal spatial adaptation by wavelet shrinkage, *Biometrika* 81, (1994), pp. 425--256

[18] R. Tibshirani, K. Knight. The covariance inflation criterion for model selection, *Journal of the Royal Statistical Society*, Ser. B 61 (1999), pp. 529-546.

[19] J. Ye. On measuring and correcting the effects of data mining and model selection, *Journal of the American Statistical Association* 93 (1998), pp. 120 -131.

[20] C.M. Hurvich, C.L. Tsai. Regression and time series model selection in small samples, *Biometrika* 76 (1989), pp. 297-307.

[21] C.M. Hurvich, C.L. Tsai. A cross-validatory AIC for hard wavelet thresholding in spatially adaptive function estimation, *Biometrika* 85, (1998), pp. 701-710.

[22] C.R. Rao, Y. Wu. A strongly consistent procedure for model selection in a regression problem, *Biometrika* 76 (1989), pp. 369-374.

[23] J. Shao. Linear model selection by cross-validation, *Journal of the American Statistical Association* 88 (1993), 486-494.

[24] C.Z. Wei. On predictive least squares principles, *The Annals of Statistics* 29 (1992), pp. 1-42.

[25] X. Zheng, W.Y. Loh. A consistent variable selection criterion for linear models with high-dimensional covariates, *Statistica Sinica* 7 (1997), pp. 311-325.

[26] Y. Benjamini, Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society*, Ser. B 57 (1995), pp. 289-300.

[27] M. Clyde, E.I. George. Empirical Bayes estimation in wavelet nonparametric regression. In: Muller, P., Vidakovic, B. (eds.) *Bayesian Inference in Wavelet-Based Models*. Springer-Verlag, New York, 1999, pp. 309-322.

[28] M. Clyde, E.I. George. Flexible empirical Bayes estimation for wavelets, *Journal of the Royal Statistical Society*, Ser. B 62 (2000), pp. 681-689.

[29] I.M. Johnstone, B.W. Silverman. Empirical Bayes approaches to mixture problems and wavelet regression. Technical Report, University of Bristol, 1998.

[30] B. Efron. Estimating the error rate of a predictive rule: improvement over cross-validation, *Journal of the American Statistical Association* 78 (1983), pp. 316-331.

[31] G. Gong. Cross-validation, the jackknife, and the bootstrap: excess error estimation in forward logistic regression, *Journal of the American Statistical Association* 393 (1986), pp. 108-113.

[32] J. Shao. An asymptotic theory for linear model selection, *Statistica Sinica* 7 (1997), pp. 229-264.

[33] P. Zhang. Inference after variable selection in linear regression models, *Biometrika* 79 (1992), pp.741-746.

[34] L. Brieman. The little bootstrap and other methods for dimensionality selection in regression: x-fixed prediction error, *Journal of the American Statistical Association* 87 (1992), pp. 738-754.

[35] L. Brieman. Better subset selection using the nonnegative garrote, *Technometrics* 37 (1995), pp. 373-384.

[36] R. Tibshirani. Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society*, Ser. B 58 (1996), pp. 267-288.

[37] E.I. George. Bayesian model selection. In: Kotz, S., Read, C., Banks, D. (eds.) *Encyclopedia of Statistical Sciences*, Update Vol. 3. Wiley, New York, 1999, pp. 39-46.

[38] P.H. Garthwaite, J.M. Dickey. Quantifying and using expert opinion for variable-selection problems in regression (with discussion), *Chemometrics and Intelligent Laboratory Systems* 35 (1996), pp. 1-34.

[39] E. Grechanovsky, I. Pinsker. Conditional p-values for the F-statistic in a forward selection procedure, *Computational Statistics & Data Analysis* 20 (1995), pp. 239-263.

[40] D.J. Hudson. *Statistics*. Geneva, 1964.