South African Statistical Journal
Vol. 57, No. 1, pp. 45–64
https://doi.org/10.37920/sasj.2023.57.1.4

Variable selection by searching for good subsets

J. H. Venter and P. J. de Jongh

Centre for Business Mathematics and Informatics, North-West University, Potchefstroom

Machine learning and statistical models are increasingly used in a prediction context and in the process of building these models the question of which variables to include often arises. Over the last 50 years a number of procedures have been proposed, especially in the statistical literature. In this paper a new variable selection procedure is introduced for linear models. A subset of variables is defined here to be "good at margin λ" if it has two properties, namely (i) its associated criterion of fit will be improved in relative terms by less than λ if any variable is added to it, and (ii) its criterion of fit will deteriorate in relative terms by at least λ if any variable inside it, is dropped from it. Thus, such a subset contains all variables that are individually important and none that are unimportant at a given margin $\lambda \geq 0$. This paper discusses calculation of such λ-good subsets. The "good" approach extends readily to generalised linear and many other models by using an appropriate criterion of performance. The approach is illustrated on an artificial data set and a number of real data sets.

Keywords: Good subsets, Linear regression, Logistic regression, Robust regression, Subset selection, Variable importance, Variable selection.

1. Introduction

The literature on variable selection methods for linear regression models is truly vast and will not be surveyed extensively here. The interested reader is referred to the recent surveys by Heinze et al. (2018), Desboulets (2018) and Talbot and Massamba (2019). At least three broad classes of methods should be mentioned. The first class is usually referred to as subset selection and consists of a variety of methods among which are best subset selection, forward, backward and stepwise selection, and combinations of these approaches (see e.g. Zhang, 2008). A second class is often referred to as "regularisation" and consists of adding a term to the criterion of fit which penalises the number or size of the parameter estimates used in the model. Among these methods are AIC (Akaike Information Criteria; see Akaike, 1992), SBC (Schwarz Bayesian Criteria; see Schwarz, 1978), ridge regression, the lasso and many further variations depending on the form of the penalisation function (see e.g. Desboulets, 2018; Freijeiro-González et al., 2022). A third class seeks to attach some measure of importance to the variables and then selects a subset from those that are most important (see e.g. Grömping, 2007; Mielniczuk and Teisseyre, 2014).

Corresponding author: P. J. de Jongh (riaan.dejongh@nwu.ac.za)
MSC2020 subject classifications: 62J12

In this paper we introduce a simple approach that has some characteristics of all three of these classes of methods. The main idea underlying selection methods is that the subset of variables selected should include all and only those variables that are "important" to establish the relationship between the regressors and the response variable. Methods diverge depending on how the notion of a variable being important (or significant, essential, etc.) is measured. In Section 2 we introduce a notion which we refer to as the selected subset being "λ-good", with λ being a tuning parameter expressing the demarcation level between a variable being important or unimportant, when importance is expressed in terms of the relative change in the measure of fit of the model. Search algorithms to calculate such subsets and their paths as functions of λ are discussed. In Section 3 these ideas are illustrated using generated data from a known linear model. Section 4 further illustrates the λ-good method applied to the Boston housing data set, which is often used in the literature to illustrate selection methods. A variety of possible choices of λ is also discussed. Section 5 illustrates the application of the method to the case of robust linear regression. Section 6 further illustrates the method applying it to a large data set and using logistic regression. Section 7 concludes with an outline of open issues and further research projects. An appendix shows formulas that ease the computational efforts required by the method.

2. The "good subsets" approach to linear regression

In this section we introduce the main ideas of the "good subsets" approach to variable selection in the simplest context, namely standard linear regression. Its extension to other cases will be discussed subsequently. To begin with some notation is needed. Suppose that we have a response (dependent or target) variable Y and K regressors (independent, explanatory or predictor variables) $X_1, X_2, \ldots, X_K$ and the model under consideration is

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_K X_K + e, \tag{1}$$

with e an error term and $\boldsymbol{b} = (b_0, b_1, \ldots, b_K)$ the regression coefficients. Not all the regressors may be relevant and some of the regression coefficients may actually be 0, but we do not know which these might be. Indeed, we are especially interested in the **sparse case** where only a few coefficients are non-zero and their identification and estimation are the main issues.

For brevity we shall refer to regressor X_k simply as regressor k below. The data consists of observations $Y_n, X_{n1}, \ldots, X_{nK}, n = 1, \ldots, N$ and the minimised error sum of squares associated with a subset $S \subseteq A = \{1, 2, \ldots, K\}$ of the regressors, is given by

$$\mathrm{ESS}(S) = \min_{\boldsymbol{b}} \sum_{n=1}^{N} \left[Y_n - b_0 - \sum_{k \in S} b_k X_{nk} \right]^2 .$$

Suppose that we have a subset S under consideration and want to examine dropping regressors from it or adding regressors to it. If a regressor k which is in S, is dropped from S, then $\mathrm{ESS}(S)$ is increased to $\mathrm{ESS}(S \setminus k)$ and, intuitively speaking, if this increase is small, then we might as well drop regressor k, but if the increase is large, then it is important to keep regressor k. The size of the increase $\delta_k(S) = \mathrm{ESS}(S \setminus k) - \mathrm{ESS}(S)$, is a possible measure of importance of regressor k in S. Next, suppose regressor k is out of S. If it is added to S, then $\mathrm{ESS}(S)$ is decreased to $\mathrm{ESS}(S \cup k)$ and if this decrease is small then we would not want to add it, but if the decrease is large, then it

is important to add regressor k. Again the size of the decrease $\delta_k(S) = \mathrm{ESS}(S) - ESS(S \cup k)$ is a possible measure of importance of regressor k when it is out of S. Note that the definition of these measures of importance $\delta_k(S)$ of regressor k with respect to the subset S, depends on whether or not regressor k is in S.

Now suppose k is in S and we decide to drop it in view its importance measured by $\delta_k(S)$. After dropping it, k is out of the smaller subset $S \setminus k$, and the importance of k with respect to $S \setminus k$ is $\delta_k(S \setminus k) = ESS(S \setminus k) - ESS(S) = \delta_k(S)$. This equation says that after dropping a regressor from a subset, its importance with respect to the remainder of the initial subset, is the same as its importance was with respect to the initial subset. This is a reasonable property of the definitions of the $\delta_k(S)$. For example, it implies that just after dropping a regressor on the evidence of its importance, you will not suddenly have different importance evidence from the reduced subset for adding it back in again. Similarly, if k is out of S, then we have the equation $\delta_k(S \cup k) = \delta_k(S)$. This implies that if you add a regressor to a subset on the evidence of its importance, you will not have different importance evidence from the enlarged subset for dropping it out again.

The importance measures $\delta_k(S)$ are not invariant with respect to the scale of the response Y. Scale invariance is desirable since it would imply that the importance measures express relationships of the regressors to the response that are more intrinsic than the unit of measurement of the response. Such scale invariance can be obtained if the increase (or decrease) in error sum of squares is expressed in relative terms. This can be done in various ways. One possibility is the measure $\delta'_k(S) = \delta_k(S)/\mathrm{ESS}(S)$. This may be viewed as a T-type statistic for testing the hypothesis that $\beta_k = 0$ (see e.g. Mielniczuk and Teisseyre, 2014). However, this measure does not satisfy the equivalents of the equations $\delta'_k(S \setminus k) = \delta'_k(S)$ and $\delta'_k(S \cup k) = \delta'_k(S)$ discussed above. An importance measure that does satisfy these equations and provides scale invariance as well, is obtained if we express the $\delta_k(S)$ relative to the geometric mean of the error sum of squares of the two subsets S involved in each case. We denote this measure by $\Delta_k(S)$. For reference purposes below, the definition is as follows:

$$\Delta_k(S) = \begin{cases} [ESS(S \setminus k) - ESS(S)]/\sqrt{ESS(S)ESS(S \setminus k)} & \text{if } k \in S, \\ [ESS(S) - ESS(S \cup k)]/\sqrt{ESS(S)ESS(S \cup k)} & \text{if } k \notin S. \end{cases} \tag{2}$$

It is straightforward to verify that $\Delta_k(S \setminus k) = \Delta_k(S)$ and $\Delta_k(S \cup k) = \Delta_k(S)$, so that the implications discussed above for the $\delta_k(S)$, continue to hold while scale invariance also holds.

Turning to the use of these measures of importance, suppose that for a given subset S, $\Delta_j(S)$ is small for all $j \notin S$, then the subset S is "good enough" in the sense that adding further individual regressors will make the model less parsimonious, without leading to appreciable immediate improvement in error sum of squares. If $\Delta_i(S)$ is large for all $i \in S$, then the subset S contains all important regressors in the sense that dropping any one of them, will lead to appreciable immediate deterioration in its error sum of squares performance. More formally, for a given $\lambda \geq 0$, we define S as "**good at margin** λ" (or λ-good for short) if it satisfies the requirement

$$\Delta_j(S) < \lambda \leq \Delta_i(S) \text{ for all } j \notin S \text{ and all } i \in S. \tag{3}$$

Thus λ represents the demarcation level below which improvement by adding regressors is no longer considered worthwhile and above which deterioration by dropping regressors is unacceptable. Another way of expressing this requirement, is to order the sequence $\Delta_1(S), \Delta_2(S), \ldots, \Delta_K(S)$

decreasingly, say as $\Delta_{l_1}(S) \geq \Delta_{l_2}(S) \geq \cdots \geq \Delta_{l_K}(S)$. Then with $|S|$ denoting the size (cardinality or number of elements) in S, we must have

$$\Delta_{l_{|S|+1}}(S) < \lambda \leq \Delta_{l_{|S|}}(S). \tag{4}$$

Note that this implies that S will then be λ-good for any value of λ within the interval $(\Delta_{l_{|S|+1}}(S), \Delta_{l_{|S|}}(S)]$. Thus for $k = 1, 2, \ldots, K$, $\{\Delta_k(S)\}$ may be thought of as a measure of the importance of the k-th regressor, associated with a given subset S and S will be λ-good if it consists of exactly those regressors with importance at least λ.

The Δs can be written in terms of the R^2-values associated with the subset S using the relation $R^2(S) = 1 - \mathrm{ESS}(S)/ESS(\emptyset)$ where $\emptyset$ is the empty set. We find

$$\Delta_k(S) = \begin{cases} \left[R^2(S) - R^2(S \setminus k)\right] / \sqrt{\left[1 - R^2(S)\right]\left[1 - R^2(S \setminus k)\right]} & \text{if } k \in S, \\ \left[R^2(S \cup k) - R^2(S)\right] / \sqrt{\left[1 - R^2(S)\right]\left[1 - R^2(S \cup k)\right]} & \text{if } k \notin S. \end{cases} \tag{5}$$

In both cases the change in the R^2-value is expressed relative to the geometric mean of the "unexplained variance fraction" associated with the two subsets in question.

In applications it is important to calculate the Δs efficiently. The appendix explains how this can be done.

Some questions arise. For a given margin λ, does there exist λ-good subsets and if so, how do we find them? What value of λ should be used for any given data set? At present we do not have complete answers to these questions but here are some relevant notes. If the empty subset $S = \emptyset$ is to be λ-good only the left-hand inequality of (3) is relevant and taking (5) into account, it requires $\lambda > R^2(j)/\sqrt{1 - R^2(j)}$ for $j = 1, 2, \ldots, K$. Thus if

$$\lambda > \lambda_{\max} = \max_{j \in A} R^2(j)/\sqrt{1 - R^2(j)}, \tag{6}$$

then the empty subset $S = \emptyset$ is λ-good. At the other end of the scale, note that (3) holds for $S = A$ if $\lambda = 0$ so that the set of all regressors is 0-good. Typically, we do not want to use all regressors for reasons of parsimony and will take $\lambda > 0$. These notes suggest the use of λ-values in the interval between 0 and the bound $\lambda_{\max}$ in (6). This bound does not depend on any unknowns such as $\mathrm{Var}(e)$, the error variance in the model (1), and this is a benefit flowing from working with the scale invariant definition of the Δs in (2).

We propose the following **search algorithm** to calculate good subsets at any given margin λ:

(1) Start with an initial subset S (e.g. the most parsimonious choice $S = \emptyset$).

(2) For $k = 1, 2, \ldots, K$ do:

- if $k \notin S$ and $\Delta_k(S) \geq \lambda$, then replace S by $S \cup k$;
- if $k \in S$ and $\Delta_k(S) < \lambda$, then replace S by $S \setminus k$.

(3) Repeat step (2) until convergence, which is obtained when a full pass through $k = 1, 2, \ldots, K$ is done without any changes to the current S.

At any stage the algorithm simply checks the Δ-value of the kth regressor associated with the current subset S against the left (right) hand side of (3) and adds k to (drops k from) S if (3) is not yet met. This algorithm is somewhat comparable to the forward-backward greedy algorithm (FoBa) introduced by Zhang (2008), but also differs in some respects. We do not evaluate the improvement (deterioration) by keeping the currently minimising regression coefficients $\boldsymbol{b}$ fixed when considering a change. Also, we do not use the greedy aspect, since it involves more computation when all possible changes are considered, to get the single best one before each change. Further we use relative (rather than absolute) size of change in ESS because this leads to scale invariance regarding the choices of λ. Finally, we work with a fixed margin rather than an adaptive threshold, which, moreover, is then the same for adding and dropping of variables, thus keeping the number of tuning parameters to one.

As explained the "good" approach to variable selection is subset selection based. Thus, coefficient estimates are either exactly zero or the least squares estimates. Compared to the regularisation (penalty) approach either no shrinking or complete shrinking to 0 takes place. The margin λ may be thought of as a (single) tuning parameter and as for the regularisation approach, a path of estimated coefficient values when λ varies can also be calculated easily. By (4) they may be taken as simple step functions of λ. We use the following good subsets **path generating algorithm**:

(0) Start with an initial choice, say $\lambda = \lambda_0$, e.g. a choice larger than the bound in (6) so that $S = \emptyset$ is good at this level and no regressors are in the model. Then calculate $\lambda_1 = \Delta_{I_{|S|+1}}(S)$ in (4).

(1) With $\lambda = \lambda_1$ apply the search algorithm above to find a λ_1-good subset S with its associated coefficient estimates. The subset of the previous step may be used as starting subset for the search algorithm here. With the new subset S, calculate $\lambda_2 = \Delta_{I_{|S|+1}}(S)$ to use in the next step.

(2) Repeat step (1) but with $\lambda = \lambda_2$ and carry on until a predefined small stopping value of λ is reached.

One possible use of the path results to make a specific choice of λ is to calculate selection criteria, such as AIC and SBC, to select a "best" choice of λ. One could also use cross-validation and other strategies for this purpose.

Many variations of the ideas set out above are possible. Instead of defining the Δs using the error sum of squares in (2) or R^2 in (5), we could work with the mean squared error (MSE) and the adjusted R^2 associated with the subset S given by

$$\mathrm{MSE}(S) = \frac{1}{N - |S| - 1}\,\mathrm{ESS}(S) \text{ and } R^2_a(S) = 1 - \frac{N-1}{N-|S|-1}\left(1 - R^2(S)\right).$$

Since $\mathrm{MSE}(S)$ is not necessarily decreasing (increasing) when an index is added to (dropped from) S, the Δ values associated with it are not necessarily non-negative, but this makes no difference to the definitions, the algorithms and their convergence stated above. We could also replace criteria based on squared error loss by absolute error and other forms of loss functions and performance criteria more appropriate for the circumstances (e.g. for robust model selection purposes or in generalised linear model contexts).

The following sections illustrate the ideas stated above. Section 3 applies it to artificial data in order to demonstrate the details of the good approach in a context where the true model and its parameters are known. The subsequent sections deal with practical data sets used in the literature.

3. Illustration in linear regression using artificial data

In this section we illustrate the essentials of the good methodology in the standard linear regression case using generated data. We took the sample size $N = 200$, used $K = 10$ regressors and generated 200 values for each of $X_1, X_2, \ldots, X_K$ taking them independent unit normally distributed and calculating the corresponding values for Y from the model (1). We took $\boldsymbol{b} = (0, 0, 0.2, 0, 0.4, 0, 0.6, 0, 1.0, 0, 0)$ and assumed e also independent unit normally distributed. Hence the true important regressors in order of increasing size of their regression coefficients are X_2, X_4, X_6 and X_8. We repeated the path generating algorithm until the remaining Δ values were below 0.001.

Table 1 summarises the main features of the method applied to this data. The top panel of Table 1 shows the details as the steps of the path generation algorithm proceed. For ease of presentation, we used the average sum of squared errors ($ASE = ESS/N$) as criterion of fit. The λ values along the path and the ASE values are shown in the second and third rows and the number of passes (count) required for convergence on each step are in the fourth row. The subset sizes are in row five of the top panel and the corresponding coefficient estimates are shown in the middle panel. The Δ values for the ASE criterion are shown in the bottom panel of Table 1. On step 0 they are all below the value $\lambda = 1$ and the empty subset is 1-good. The largest Δ value is 0.4569 (that of X_8) and this is the λ value taken for step 1. On step 1 the Δ of X_8 stays at 0.4569 but the Δ values of all other regressors are smaller than 0.4569 so that selecting only X_8 at this stage yields a 0.4569-good subset. The largest Δ value among the other regressors is 0.3226 (that of X_6). Hence the subset consisting of only X_8 is λ-good for any λ in the interval $(0.3226, 0.4569]$. For step 2 the algorithm takes $\lambda = 0.3226$ and then selects the subset consisting of X_8 and X_6. This yields a λ-good subset for any λ in the interval $(0.1042, 0.3226]$ since 0.1042 is the largest Δ value among the other regressors not in this subset. Step 3 then carries on with $\lambda = 0.1042$ and adds X_4. Step 4 proceeds in the same manner, adding X_2 to get a subset which is a λ-good subset for any λ in the interval $(0.0187, 0.0342]$. The coefficient estimates in the middle panel show that at step 4 all four known important regressors were identified and included in the selected subset and that the estimates are quite in line with the true coefficient values. One more regressor was added at each of steps 5 to 8, at which stage the selected subset consisted of all but the regressors X_1 and X_7. Their Δ values were below the level of 0.001 and the algorithm stopped. The coefficient estimates as functions of λ are shown graphically in the lefthand panel of Figure 1, making it clear that they form simple step functions here.

Rows 6 to 8 of the top panel in Table 1 show the R^2, AIC and SBC criteria of the subsets at each step. The R^2 values increase but are quite constant from step 4 onwards. AIC has a shallow minimum at step 6 and the SBC has a somewhat clearer minimum at step 4. These indications are in line with the true model on which this data was based.

In the exposition of the λ-good method in Section 2, no distributional assumptions were made regarding the error terms e_n. If we assume that they are independent and identically distributed (iid) with zero expectation and variance σ^2, then an estimate of σ^2 corresponding the selected subset S are given by $\widehat{\sigma}^2(S) = ESS(S)/(N - 1 - |S|)$. Row 9 of Table 1 shows the values of $\widehat{\sigma}$ for each subset. The right-hand panel of Figure 1 plots these estimates as functions of λ. This graph is also a step function since the selected subset do not change when λ varies over the successive intervals.

Table 1. Illustration of path features of good subsets algorithm using ASE as criterion applied to the generated data.

Details on each step									
step	0	1	2	3	4	5	6	7	8
λ	1.0000	0.4569	0.3226	0.1042	0.0342	0.0189	0.0134	0.0026	0.0022
ASE	2.1416	1.3614	0.9874	0.8897	0.8598	0.8437	0.8325	0.8303	0.8285
count	1	2	2	2	2	2	2	2	2
size	0	1	2	3	4	5	6	7	8
R-square	0.0000	0.3643	0.5389	0.5846	0.5985	0.6060	0.6113	0.6123	0.6131
AIC	356.31	267.71	205.46	186.62	181.79	180.02	179.35	180.82	182.38
SBC	157.61	72.31	13.36	−2.18	−3.72	−2.19	0.43	5.20	10.06
$\widehat{\sigma}$	1.4671	1.1727	1.0012	0.9528	0.9391	0.9327	0.9288	0.9300	0.9314
Regressor coefficient estimates on each step									
intcpt	-0.0522	0.0138	-0.0119	-0.0055	0.0155	0.0036	-0.0007	-0.0033	-0.0019
X1	0	0	0	0	0	0	0	0	0
X2	0	0	0	0	0.1718	0.1730	0.1689	0.1641	0.1665
X3	0	0	0	0	0	-0.1322	-0.1228	-0.1256	-0.1304
X4	0	0	0	0.3431	0.3440	0.3418	0.3545	0.3503	0.3491
X5	0	0	0	0	0	0	0.1041	0.1061	0.1054
X6	0	0	0.6622	0.6699	0.6538	0.6548	0.6628	0.6631	0.6571
X7	0	0	0	0	0	0	0	0	0
X8	0	0.8585	0.9701	0.9871	0.9764	0.9811	0.9907	0.9885	0.9839
X9	0	0	0	0	0	0	0	0	-0.0408
X10	0	0	0	0	0	0	0	-0.0474	-0.0464
Deltas on each step when using ASE as criterion									
X1	0.0013	0.0028	0.0002	0.0001	0.0000	0.0001	0.0002	0.0001	0.0002
X2	0.0334	0.0373	0.0302	0.0342	0.0342	0.0353	0.0340	0.0319	0.0328
X3	0.0039	0.0115	0.0173	0.0178	0.0189	0.0189	0.0164	0.0171	0.0183
X4	0.0302	0.0677	0.1042	0.1042	0.1082	0.1088	0.1168	0.1137	0.1131
X5	0.0001	0.0017	0.0079	0.0171	0.0158	0.0134	0.0134	0.0139	0.0137
X6	0.0939	0.3226	0.3226	0.3596	0.3529	0.3596	0.3698	0.3709	0.3607
X7	0.0136	0.0097	0.0012	0.0001	0.0011	0.0005	0.0001	0.0001	0.0001
X8	0.4569	0.4569	0.6935	0.7674	0.7722	0.7877	0.8020	0.7994	0.7887
X9	0.0151	0.0081	0.0009	0.0006	0.0012	0.0024	0.0023	0.0022	0.0022
X10	0.0080	0.0048	0.0062	0.0030	0.0014	0.0021	0.0026	0.0026	0.0025

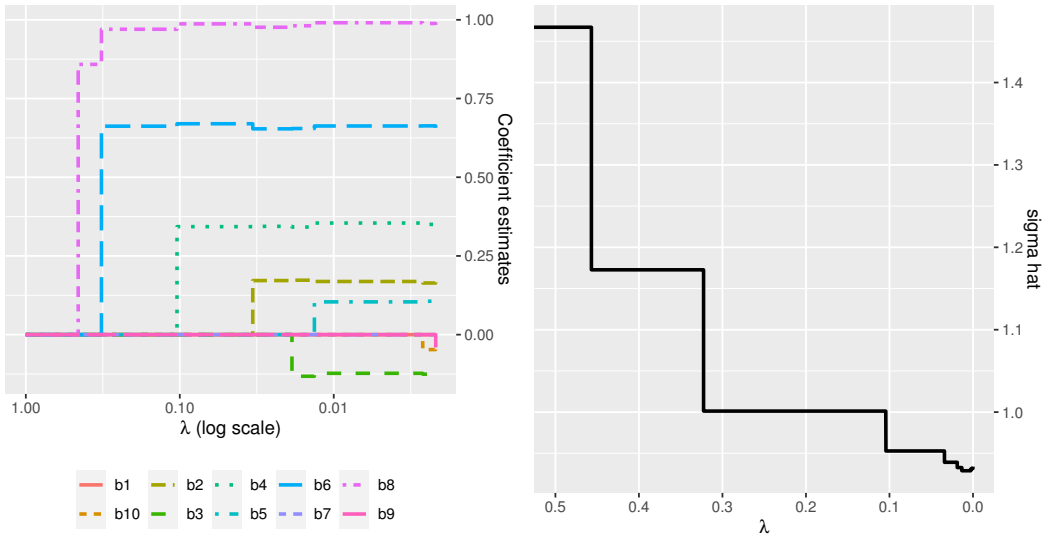

Figure 1. Path of estimated regression coefficients (left panel) and standard deviation estimate as functions of λ.

4. Illustration in standard regression using the Boston housing data

The well-known Boston housing data set is used this section. This data set has been used for illustration purposes in many papers (e.g. Zhang, 2008). It has 506 observations with housing price as response variable and 13 other regressors as listed in Table 2 . We used Proc IML of SAS to program the algorithms together with the SUBMIT call to Proc GENMOD to compute the model fitting details and criteria.

Path results

Table 2 shows the results in similar form to Table 1 above. Again, on step 0 the empty subset is 1-good. The largest Δ value on this step is 0.8059 (that of "lstat") and this is the λ value taken for step 1. Selecting only lstat yields 0.8059-good subset. The largest Δ value among the other regressors is 0.2326, that of "rooms". Hence selecting only lstat is λ-good for any λ in the interval $(0.2326, 0.8059]$. For step 2 the algorithm takes $\lambda = 0.2326$ and then selects both lstat and rooms as variables to include and go on to the subsequent steps. The coefficient estimates in the middle panel show that one more regressor was added at step 3, two more at step 4, then again one more on steps 5, 6 and 7, three more on step 8 and one more on step 9. At that stage the selected subset consisted of all but the regressor "age" whose Δ value was zero (to four decimals) and the algorithm stopped.

We calculated the ASEs of all subsets of given sizes separately, compared them to those of the good subsets in Table 2 and found that the good subsets were in all cases actually the best of their sizes, i.e. they possessed the smallest ASEs among all subsets of their respective sizes. While this is true for this data set (and some others we looked at), we doubt that it would be generally true, and this note leads to the issue of finding conditions under which the "good" subsets are guaranteed to

Table 2. Illustration of path features of good subsets algorithm using ASE as criterion applied to the Boston housing data.

Details on each step										
step	0	1	2	3	4	5	6	7	8	9
λ	1.0000	0.8059	0.2326	0.1175	0.0370	0.0267	0.0227	0.0161	0.0081	0.0002
ASE	84.42	38.48	30.51	27.13	24.64	23.99	23.46	23.08	21.90	21.89
count	1	2	2	2	3	2	2	2	4	2
size	0	1	2	3	5	6	7	8	11	12
R^2	0.0000	0.5441	0.6386	0.6786	0.7081	0.7158	0.7222	0.7266	0.7406	0.7406
AIC	2755	2359	2244	2186	2141	2130	2120	2114	2094	2096
SBC	2251	1859	1748	1695	1659	1652	1646	1644	1636	1643
Regressor coefficient estimates on each step										
intcpt	22.533	34.554	-1.358	18.567	37.499	36.923	30.412	30.317	36.341	36.437
crim	0	0	0	0	0	0	0	0	-0.1084	-0.1080
zn	0	0	0	0	0	0	0	0.0378	0.0458	0.0463
indus	0	0	0	0	0	0	0	0	0	0.0206
chas	0	0	0	0	0	3.2443	3.0519	3.1111	2.7187	2.6890
nox	0	0	0	0	-17.997	-18.740	-16.677	-16.687	-17.376	-17.714
rooms	0	0	5.0948	4.5154	4.1633	4.1118	4.2944	4.1161	3.8016	3.8144
age	0	0	0	0	0	0	0	0	0	0
distance	0	0	0	0	-1.1847	-1.1446	-1.1235	-1.3827	-1.4927	-1.4786
radial	0	0	0	0	0	0	0	0	0.2996	0.3058
tax	0	0	0	0	0	0	0	0	-0.0118	-0.0123
pt	0	0	0	-0.9307	-1.0458	-1.0027	-0.9737	-0.8819	-0.9465	-0.9522
b	0	0	0	0	0	0	0.0090	0.0094	0.0093	0.0093
lstat	0	-0.9500	-0.6424	-0.5718	-0.5811	-0.5698	-0.5372	-0.5431	-0.5226	-0.5239
Deltas on each step when using ASE as criterion										
crim	0.1636	0.0076	0.0204	0.0090	0.0114	0.0096	0.0047	0.0081	0.0081	0.0217
zn	0.1393	0.0083	0.0037	0.0011	0.0122	0.0136	0.0161	0.0161	0.0230	0.0232
indus	0.2674	0.0051	0.0040	0.0001	0.0014	0.0022	0.0013	0.0014	0.0002	0.0002
chas	0.0312	0.0412	0.0362	0.0279	0.0267	0.0267	0.0241	0.0254	0.0203	0.0197
nox	0.2020	0.0002	0.0010	0.0018	0.0591	0.0654	0.0515	0.0524	0.0477	0.0459
rooms	0.6728	0.2326	0.2326	0.2024	0.1861	0.1862	0.2024	0.1861	0.1633	0.1630
age	0.1534	0.0157	0.0013	0.0048	0.0000	0.0002	0.0008	0.0001	0.0000	0.0000
distance	0.0645	0.0405	0.0230	0.0370	0.0944	0.0903	0.0889	0.1038	0.1230	0.1152
radial	0.1576	0.0013	0.0118	0.0004	0.0043	0.0048	0.0122	0.0080	0.0442	0.0425
tax	0.2485	0.0142	0.0279	0.0032	0.0008	0.0003	0.0002	0.0003	0.0244	0.0216
pt	0.2993	0.1476	0.1175	0.1175	0.1569	0.1473	0.1416	0.1106	0.1034	0.1029
b	0.1179	0.0102	0.0337	0.0288	0.0253	0.0227	0.0227	0.0252	0.0241	0.0243
lstat	0.8059	0.8059	0.3588	0.3126	0.2583	0.2546	0.2253	0.2327	0.2202	0.2199

be the "best" subsets. This issue is an open problem at present. Note also that the results potentially depend on the order in which the regressors are presented. To check on whether this is the case here, we randomly permuted the sequencing of the regressors in the data set and presented the permuted data to the algorithm but found that this made no difference to the results. Again, it is an open issue to what extent this is generally true.

Further, we calculated the R^2, AIC and SBC values along the path and show them in rows 6 to 8 of Table 2. R^2 increases with model size but only slowly towards the end; both AIC and SBC achieve their minima on step 8, suggesting a choice of λ in the interval $[0.0081, 0.0002)$ and the corresponding model with 11 regressors shown in the table.

Other choices of λ

Next, we discuss two other methods to choose λ. The first one is based on data-splitting and cross-validation ideas. Some notation is needed to formulate this clearly. Let $\mathcal{N} = \{1, 2, \ldots, N\}$ denote the indices of the observations and let $\mathcal{D} \subset \mathcal{N}$ denote a subset consisting of $N/2$ indices when N is even and one more otherwise. Cross-validation fits a model on the part of the data in $\mathcal{D}$ (or $\mathcal{D}^c = \mathcal{N} \setminus \mathcal{D}$) and uses the fitted model to judge how well it predicts the responses in $\mathcal{D}^c$ (or $\mathcal{D}$). More formally, for a given subset of regressors S define the average total squared error when fitting separate models to the two data parts by

$$\mathrm{ASE}(S, \mathcal{D}) = \frac{1}{N}\left[\min_{\boldsymbol{b}} \sum_{n\in\mathcal{D}} \left[Y_n - b_0 - \sum_{k\in S} b_k X_{nk}\right]^2 + \min_{\boldsymbol{b}} \sum_{n\in\mathcal{D}^c} \left[Y_n - b_0 - \sum_{k\in S} b_k X_{nk}\right]^2\right]. \quad (7)$$

With $\boldsymbol{b}^{\mathcal{D}} = (b_0^{\mathcal{D}}, b_1^{\mathcal{D}}, \ldots, b_K^{\mathcal{D}})$ and $\boldsymbol{b}^{\mathcal{D}^c} = (b_0^{\mathcal{D}^c}, b_1^{\mathcal{D}^c}, \ldots, b_K^{\mathcal{D}^c})$ denoting the choices of $\boldsymbol{b}$ at which the two minima in (7) are achieved, define the average squared error when cross-predicting between the data parts by

$$\mathrm{APSE}(S, \mathcal{D}) = \frac{1}{N}\left[\sum_{n\in\mathcal{D}} \left[Y_n - b_0^{\mathcal{D}^c} - \sum_{k\in S} b_k^{\mathcal{D}^c} X_{nk}\right]^2 + \sum_{n\in\mathcal{D}^c} \left[Y_n - b_0^{\mathcal{D}} - \sum_{k\in S} b_k^{\mathcal{D}} X_{nk}\right]^2\right].$$

Then take the Δs as in (2) with $ASE(S)$ replaced by $ASE(S, \mathcal{D})$, apply the path generating algorithm to obtain the sequence of λ value intervals and associated subsets S and also calculate their $\mathrm{APSE}(S, \mathcal{D})$ values. Then a choice $\lambda = \lambda_{\min}(\mathcal{D})$ that is in the interval minimising $\mathrm{APSE}(S, \mathcal{D})$ along the path is reasonable from a cross-validation point of view. This still depends on the choice of subset $\mathcal{D}$ and the question now is what to do about this. One possibility is to try to match the data items in $\mathcal{D}$ and $\mathcal{D}^c$ as closely as possible, but this will not be pursued here. The more usual way is to take many random choices and average the $\lambda_{\min}(\mathcal{D})$ over repetitions. Doing this 1000 times delivered the average value of 0.0019 when we took the $\lambda_{\min}(\mathcal{D})$ at the midpoint of the minimising intervals in each case. The value 0.0019 for λ is in the interval of step 7 in Table 2 suggesting a somewhat more parsimonious model than that suggested by AIC and SBC, involving only eight regressors rather than 11 . We also calculated the average of the minimised $\mathrm{APSE}(S, \mathcal{D})$ over repetitions and found that this was 24.50 which is only about 6% above the ASE of 23.08 in Table 2 on step 7. This suggests that the ASE of 23.08 for the chosen model is only slightly overoptimistic in terms of error sizes that may be expected when predicting out-of-sample with it.

Another method to choose λ can be based on the addition to the data of pseudoregressors which are known to be unrelated to the response. Adapting from Wu et al. (2007), we randomly permuted the rows of the regressor data and then added them to the existing data as new (or pseudo-) regressors. The random permutation destroys the relationship between the response and the regressors so that these pseudoregressors should not be included in the selected model. A table similar to Table 2 showed that when the good approach is applied to this extended data set, everything stayed the same up to step 7 with the coefficient estimates of the pseudoregressors all having the value zero. On step 8 one of the pseudoregressors (nox) was included in the selected model with a non-zero coefficient. So, at this point a regressor that is known to be unrelated to the response (at least to the extent to which the randomisation destroyed their relationship), is selected ahead of the remaining actual regressors. This suggests that the remaining unselected regressors are also likely to be unrelated to the response and that we could therefore stop at the previous step, i.e. choose λ in $[0.0160, 0.0081)$. But this was just one randomised addition of pseudoregressors. We repeated the process independently 1000 times and averaged the resulting λ-values, getting 0.0090. This value is also in step 8 of Table 2 and therefore agrees with the choice found by the cross-validation method above. There are many other ways to generate pseudoregressors. Taking them to be independent standard normally distributed leads to virtually the same results. There are some further issues with this method of choosing λ. We took their number equal to the number of actual regressors, but it is not clear that this is necessary or desirable, especially when we already have a large number of regressors, where this will double their number and add to the computational burden. Our application of the pseudoregressors method is rather different from that given in Wu et al. (2007) and it is evident that further research is required to establish the properties of our adaptation of this idea.

Adding interactions

To see what happens when the number of regressors are larger, we added all interactions to the main regressors, thus increasing the number of regressors to 91 . Table 3 shows the results when the good approach is applied to the data with this much extended set of variables. Table 3 is similar in form to Table 2 but leaves out the coefficient estimates of the variables that did not enter up to step 9. Also, all the Δs were left out since their use is as before.

Among notable features when comparing Tables 2 and 3 are the following. The three main regressors rooms, pt and lstat which entered early in Table 2 also enter early in Table 3 and moreover figure prominently in many of the interactions in the selected subsets. This suggests that they are indeed relevant to modeling this data. Main regressors such as crim, indus, age, radial and tax that entered late in Table 2, tend to enter late both as main regressors and as interactions. This suggests that they are relatively irrelevant. The R^2 criterion increases faster in Table 3 as the steps proceed. For example, on step 5 the selected subsets in both tables contained 6 variables, while the R^2 values were 0.792 and 0.716 respectively. This implies that the good approach did indeed find better variables to select when the interactions were allowed. The AIC and SBC criteria continued to decline as the steps progressed, suggesting that from the point of view of these criteria, even better models are beyond those in Table 3 . However, to go beyond the steps in the tables imply using very small λ-values which may be a case of scraping the bottom of the barrel and leading to overfitting.

Table 3. Illustration of path features of good subsets algorithm using ASE as criterion applied to the Boston housing data with interactions added as variables.

Details on each step										
step	0	1	2	3	4	5	6	7	8	9
λ	1.000	0.852	0.241	0.080	0.074	0.065	0.063	0.023	0.020	0.014
ASE	84.42	36.88	29.00	20.20	18.76	17.58	15.48	12.60	11.84	11.68
count	1	2	2	4	2	2	2	4	2	2
size	0	1	2	4	5	6	8	12	14	15
R^2	0.000	0.563	0.657	0.761	0.778	0.792	0.817	0.851	0.860	0.862
AIC	2755	2338	2218	2039	2004	1973	1912	1816	1789	1784
SBC	2251	1838	1722	1552	1521	1494	1442	1363	1344	1343
Regressor coefficient estimates on each step										
intcpt	22.53	33.86	-1.07	-22.82	-17.25	-91.47	-110.3	-158.0	-200.8	-205.0
nox	0	0	0	0	0	0	0	0	129.9	134.0
rooms	0	0	5.006	10.54	10.05	21.46	24.07	31.63	35.37	35.70
pt	0	0	0	0	0	4.402	5.239	7.596	7.745	7.969
lstat	0	0	0	1.857	1.720	1.290	1.539	3.014	1.555	1.429
crim*chas	0	0	0	0	0	0	1.316	1.098	1.162	1.066
nox*room	0	0	0	0	0	0	0	-3.747	-15.50	-15.68
crim*dist	0	0	0	0	0	0	0	-0.094	-0.098	-0.212
indu*radi	0	0	0	0	0	0	0	0.028	0.037	0.044
dist*tax	0	0	0	0	-0.002	-0.002	-0.002	-0.003	-0.003	-0.003
nox*pt	0	0	0	0	0	0	0	0	-3.196	-3.374
room*pt	0	0	0	-0.107	-0.099	-0.781	-0.904	-1.188	-1.049	-1.068
crim*lstat	0	0	0	0	0	0	-0.006	0	0	0.009
indu*lstat	0	0	0	0	0	0	0	0	-0.011	-0.013
room*lstat	0	0	0	-0.419	-0.406	-0.335	-0.359	-0.337	-0.277	-0.248
radil*lstat	0	0	0	0	0	0	0	-0.017	-0.024	-0.030
pt*lstat	0	-0.047	-0.033	0	0	0	0	-0.074	0	0

Remarks

We also carried out the above calculations using MSE (rather than ASE) as fitting criterion and found that while the λ-values along the path were slightly different from those in the tables above, the selected models and their coefficient estimates were the same in the non-interactions case and closely similar in the interactions case.

5. Robust regression illustration

Using only ASE or MSE as measures of performance in regression do not allow for issues of possible outliers and lack of robustness. To illustrate that the good subsets approach can be applied with only little adaptation in robust contexts, we use the college data set AER in the R package. This data set is also used by Dupuis and Victoria-Feser (2013) to illustrate their robust version of the VIF-regression forward subset selection method of Lin et al. (2011). It has 4739 observations and 14 regressors. Following Dupuis and Victoria-Feser (2013), we have standardised the regressors data in the results below.

There are presently many possible criteria of fit that do take robustness issues into account. To make our results comparable to those of Dupuis and Victoria-Feser (2013), we use the weighted least squares criterion of M-estimation as implemented in SAS Proc Robustreg, taking for the weight function the Tukey biweights with constant $c = 4.685$. For a given subset S, Robustreg produces a robust version of the ASE which we can use to calculate the corresponding Δs and the rest of the good subsets method then proceeds as before. Table 4 shows the results in the same form as the previous tables. The robust R^2 quickly rises, achieving the value 0.3658 on step 4 and thereafter only rises slowly to 0.3706 on step 7. Both the robust AIC values achieve their minimum on step 7, suggesting a model with 9 regressors, while SBC flags a model with one less regressors on step 6. The middle panel of Table 4 shows the coefficient estimates and, except for including the regressor gender and excluding urban, these agree well with those of robVIF in Table 3 of Dupuis and Victoria-Feser (2013). The bottom panel of Table 4 shows the observation numbers identified as outliers by RobustReg along the λ-path. Except for number 4515, once an outlier is flagged it stays flagged along the path, so that the choice of λ is not very important in this regard.

To apply the cross-validation method of choosing λ, requires some adaptation to take the possibility of outliers into account. One way to do it is to find the outliers identified in each data part by Robustreg when fitting the model to that part and then to ignore these outliers when calculating the APSE from the cross-predictions. The average choice of λ minimising this adapted APSE is then calculated as before. The pseudoregressors method do not require adaptation for possible outliers and can be applied as before. We found that the average λ choices according to the two methods were 0.0028 and 0.0010 which both lie in the interval of step 6 of Table 4 and therefore substantially agree with the choice suggested by AIC and SBC in this case.

Overall, the λ-good approach to variable selection applies readily in robust regression contexts and needs only little adaptation and programming. We found this also to be the case when dealing with generalised linear models, as illustrated in the next section.

Table 4. Illustration of path features of good subsets algorithm using robust ASE as criterion applied to the college data.

				Details on each step				
step	0	1	2	3	4	5	6	7
λ	1.0000	0.3653	0.0532	0.0113	0.0068	0.0035	0.0034	0.0006
cnt	1	2	2	2	3	2	2	2
size	0	1	2	3	6	7	8	9
R^2	0.0000	0.3047	0.3407	0.3481	0.3658	0.3680	0.3702	0.3706
AIC	9892	8172	7922	7870	7746	7731	7717	7716
SBC	5157	3444	3200	3155	3050	3042	3034	3040
				Regressor coefficient estimates on each step				
intcp	13.77	13.77	13.77	13.76	13.76	13.76	13.76	13.76
gender	0	0	0	0	0	0.0680	0.0671	0.0651
afam	0	0	0	0	0.1460	0.1459	0.1370	0.1382
hisp	0	0	0	0	0.1360	0.1384	0.1389	0.1361
score	0	0.9006	0.8097	0.7952	0.8379	0.8426	0.8387	0.8373
fcoll	0	0	0.3440	0.2758	0.2377	0.2383	0.2338	0.2369
mcoll	0	0	0	0.1659	0.1453	0.1443	0.1428	0.1445
home	0	0	0	0	0	0	0	0
urban	0	0	0	0	0	0	0	0
unemp	0	0	0	0	0	0	0	0.0578
wage	0	0	0	0	0	0	0	0
distance	0	0	0	0	0	0	-0.0647	-0.0809
tuition	0	0	0	0	0	0	0	0
income	0	0	0	0	0.1841	0.1879	0.1844	0.1855
region	0	0	0	0	0	0	0	0
Obs No				**Outlier indicators on each step**				
1614	0	0	0	0	0	1	1	1
1649	0	1	1	1	1	1	1	1
1976	0	0	0	0	0	1	1	1
2161	0	0	0	1	1	1	1	1
2963	0	1	1	1	1	1	1	1
3107	0	1	1	1	1	1	1	1
4194	0	0	0	0	0	0	1	1
4515	0	0	1	1	0	0	0	0
4594	0	1	1	1	1	1	1	1
4711	0	1	1	1	1	1	1	1

6. Logistic regression illustration

In this section we illustrate the λ-good approach when dealing with a large data set and a response variable Y taking only two values, namely 0 or 1. The standard model in this context is logistic regression and we base fitting on the log-likelihood function. Some additional notation is required. For a given subset S of the regressors, put

$$P_n(S) = 1/\left\{1 + \exp\left(-b_0 - \sum_{k\in S} b_k X_{nk}\right)\right\} \tag{8}$$

and denote the negative of twice the weighted log-likelihood function corresponding to S by

$$M2LL(S) = -2\sum_{n=1}^{N} w_n \left[Y_n \log\left(P_n(S)\right) + (1 - Y_n)\log\left(1 - P_n(S)\right)\right].$$

Here, w_n is the weight assigned to the nth observation, more details of which are given below. $M2LL(S)$ is the fitting criterion that now replaces the SSE and other criteria used above. In this case, scale invariance is no longer of concern and we define the $\Delta_k(S)$ simply as the differences

$$\Delta_k(S) = \begin{cases} M2LL(S \setminus k) - M2LL(S) & \text{if } k \in S, \\ M2LL(S) - M2LL(S \cup k) & \text{if } k \notin S. \end{cases}$$

The rest of the λ-good approach then operates as before. For illustration purposes, the credit card dataset available on the website of Kaggle (see https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud) is used. The contents and context of this dataset are described in that reference. In brief, it contains data on $N = 284807$ transactions of which 492 are fraudulent ($Y_n = 1$) and the remainder clean ($Y_n = 0$). There are $K = 30$ regressors of which 28 are PCA transformations of features that are not detailed due to confidentiality issues; the other two regressors are described as 'Time' and 'Amount'. Here we refer to the regressors simply as $X_1, X_2, \ldots, X_{30}$. This is a large dataset and the illustration below follows the modelling paradigm of dividing the dataset into training and testing sets. The rows of the dataset were permutated randomly, and then half of them were put into the training set and the remainder into the testing set. Thus, the training dataset has $N_{\text{train}} = 142404$ observations with $N_{\text{train},1} = 245$ frauds and $N_{\text{train},0} = 142159$ cleans. This dataset is highly unbalanced since the fraction of frauds is only 0.172% of the total. The assignments of the weights in the fitting criterion are used to address this matter. We take $w_n = 1/2N_{\text{train},1}$ if $Y_n = 1$ and $w_n = 1/2N_{\text{train},0}$ if $Y_n = 0$. Then the totals of the weights for the frauds and the cleans are $1/2$ each and the overall total of all the weights is 1 . This enables the small number of frauds to play a meaningful role in the model training. Table 5 presents the results.

The λ used on each step is in the second row. The third row shows the values of the present fitting criterion M2LL, decreasing as more variables are added. Rows four and five show the number of passes in the search required for convergence and the subset sizes on each step.

Instead of recording criteria such as the AIC and SBC of the previous tables, rows six to nine show various fractions of correct classification for the models on each step. Row six shows the fraction of correct classifications of the frauds in the training data when the fitted model is applied. Here the probability of a fraud for each transaction is calculated using (8) with the estimated regression

Table 5. Illustration of path features of good subsets algorithm using logistic regression with fitting criterion based on weighted log-likelihood applied to the credit card fraud data.

Details on each step										
step	0	1	2	3	4	5	6	7	8	9
λ	1	0.9189	0.1399	0.0242	0.0159	0.0139	0.0068	0.0050	0.0048	0.0046
M2LL	1.386	0.4674	0.3275	0.3033	0.2873	0.2734	0.2595	0.2545	0.2497	0.2451
count	1	2	2	2	2	3	2	2	2	2
size	0	1	2	3	4	5	7	8	9	10
$cortrn1$	0	0.8645	0.9124	0.9044	0.9124	0.9163	0.9243	0.9283	0.9323	0.9323
$cortrn0$	1	0.9728	0.9657	0.9710	0.9706	0.9699	0.9720	0.9698	0.9716	0.9745
$cortst1$	0	0.8672	0.8797	0.8880	0.8963	0.8921	0.8963	0.8963	0.9004	0.8921
$cortst0$	1	0.9725	0.9667	0.9722	0.9713	0.9704	0.9723	0.9704	0.9720	0.9745
Regressor coefficient estimates on each step										
intcpt	0	-2.174	-3.012	-2.984	-3.143	-3.275	-3.463	-3.570	-3.649	-3.647
x1	0	0	0	0	0	0	0	0	0	0
x2	0	0	0	0	0	0	0	0	0	0
x3	0	0	0	0	0	0	0	0	0	0
x4	0	0	0.8614	0.7504	0.7628	0.8052	0.9193	0.9346	0.9395	0.9386
x5	0	0	0	0	0	0	0	0	0	0
x6	0	0	0	0	0	0	0	0	-0.2735	-0.2977
x7	0	0	0	0	0	0	0	0	0	0
x8	0	0	0	0	-0.3918	-0.4188	-0.4610	-0.4507	-0.5107	-0.5309
x9	0	0	0	0	0	0	0	0	0	0
x10	0	0	0	0	0	0	-0.3845	-0.4264	-0.4180	-0.4232
x11	0	0	0	0	0	0	0	0	0	0.3535
x12	0	0	0	-0.660	-0.7104	-0.6737	-0.6678	-0.7002	-0.6977	-0.6660
x13	0	0	0	0	0	0	-0.4223	-0.4131	-0.4030	-0.4044
x14	0	-1.167	-1.0765	-0.944	-0.9472	-1.0328	-0.9084	-0.9199	-0.9270	-0.8680
x15-21	0	0	0	0	0	0	0	0	0	0
X22	0	0	0	0	0	0	0	0.5470	0.6264	0.6468
X23	0	0	0	0	0	-0.3649	-0.3582	-0.4086	-0.3498	-0.3095
X24-30	0	0	0	0	0	0	0	0	0	0

coefficients. The transaction is classified as fraudulent if this estimated probability is above 0.5 and as clean otherwise. Other than for step 0 which uses no regressors, the correct classification rates are about 90% if the best two or more regressors are used. Row eight shows the same rates for the testing data. Note that these rates are just slightly lower than those for the training data on which the models are based. Hence, the models performed quite well when applied out of the training sample. Rows seven and nine show the correct prediction rates of the clean transactions in the training and testing data. These rates are in the region of 97% when using two or more regressors. That these are higher than those of the frauds, are due to the small proportion of frauds in the data. However, the choice of the weights used here is important and beneficial to identifying the frauds. For example, if all observations in the training data were given the same weight, regardless of whether or not they were frauds, then it tumed out that the percentage of correct classification of frauds, is typically below 60% while that of cleans is over 99%. Hence the use of the adjusted weights greatly improves the identification of the extremely rare frauds in the data.

As before, the bottom panel of Table 5 shows the regression coefficient estimates. It is clear that using a parsimonious model with only the four variables X_{14}, X_4, X_{12} and X_8 would accomplish fraud identification with quite high certainty and excellent generalisation properties, at least in terms of the criteria used here. Of course, there are many other criteria in logistic regression that can also be used. Among these are the "area under the curve" (AUC) and others. The λ-good method can also be formulated and carried out with such criteria.

7. Conclusion

In this paper we presented a fresh approach to linear regressor subset selection, namely finding subsets consisting of all those regressors so important that dropping any one of them deteriorates the fit quality of the model below a specified demarcation level, while simultaneously containing no regressors without this property. The demarcation level constitutes the single tuning parameter λ of the approach and such a subset is termed to be λ-good. Using a number of examples, we demonstrated that this selection method readily applies to standard linear regression and also to robust regression models. We can also report that it worked well in generalised linear models such as logistic regression and other examples of which we do not include all the detailed results here.

A number of open issues requiring further research were noted in the text above. Among these are the possible relation between "good" subsets and "best" subsets and our adaptation and application of the ideas of cross-validation and pseudoregressors to choose the tuning parameter λ. Moreover, convergence issues are always pertinent when dealing with search algorithms. In all the illustrations we found convergence within at most four passes through the list of variables (and often only two). We never found cases of non-convergence. Still, it would be useful to get theoretical confirmation that this is generally true. As with the choice of tuning parameters in all subset selection methods, more can also be done to pin down the choice of λ when dealing with practical data. Perhaps the most important outstanding issue at this stage, is to carry out a simulation based systematic comparison between the λ-good approach and other existing variable selection methods. Future research to this effect is in progress.

Acknowledgements. This work is based on research supported in part by the Department of Science and Innovation (DSI) of South Africa. The grant holder acknowledges that opinions, findings and

conclusions or recommendations expressed in any publication generated by DSI-supported research are those of the authors and that the DSI accepts no liability whatsoever in this regard.

Appendix: Calculation of the Δs for linear models

Consider an $M \times M$ real positive definite symmetric matrix $\boldsymbol{A}$ partitioned as

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{a}_{12} \\ \boldsymbol{a}'_{12} & a_{22} \end{bmatrix}, \tag{A.1}$$

with $\boldsymbol{A}_{11}$ an $(M-1) \times (M-1)$ submatrix. Then we have

$$\boldsymbol{A}^{-1} = \boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_{11} & \boldsymbol{d}_{12} \\ \boldsymbol{d}'_{12} & d_{22} \end{bmatrix}, \tag{A.2}$$

where

$$\begin{gathered} \boldsymbol{D}_{11} = \boldsymbol{A}_{11}^{-1} + d_{22}\boldsymbol{A}_{11}^{-1}\boldsymbol{a}_{12}\boldsymbol{a}'_{12}\boldsymbol{A}_{11}^{-1}, \boldsymbol{d}_{12} = -d_{22}\boldsymbol{A}_{11}^{-1}\boldsymbol{a}_{12}, \\ d_{22} = 1 \Big/ \left(a_{22} - \boldsymbol{a}'_{12}\boldsymbol{A}_{11}^{-1}\boldsymbol{a}_{12}\right) \text{ and } \boldsymbol{A}_{11}^{-1} = \boldsymbol{D}_{11} - \boldsymbol{d}_{12}\boldsymbol{d}'_{12}/d_{22}. \end{gathered} \tag{A.3}$$

This follows from simple matrix calculations but is also a special case of Muirhead (1982), Theorem A5.2. We apply these results to least squares calculations in the context of calculating the Δ_k(S).

In matrix form the linear model in Section 2 may be written as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$ where $\boldsymbol{Y}$ and $\boldsymbol{e}$ are the $N \times 1$ vectors with nth components Y_n and e_n respectively, $\boldsymbol{X}$ is the matrix given by $\boldsymbol{X} = [\mathbf{1}\, \boldsymbol{X}_1\, \boldsymbol{X}_2\, \ldots\, \boldsymbol{X}_K]$, where $\mathbf{1}$ is the $N \times 1$ vector with all components equal to 1 and, for $k = 1, 2, \ldots, K$, $\boldsymbol{X}_k$ is the $N \times 1$ vector with nth component X_{nk} while $\boldsymbol{b}$ is the $(K+1) \times 1$ vector with first component b_0 and $(k+1)$th component b_k. For a given subset $S \subseteq A = \{1, 2, \ldots, K\}$, write $\boldsymbol{X}(S)$ for the sub-matrix with first column equal to $\mathbf{1}$ and the remainder those corresponding to S, i.e. $\boldsymbol{X}(S) = [\mathbf{1}\ \{\boldsymbol{X}_k, k \in S\}]$. Similarly, denote by $\boldsymbol{b}(S)$ the sub-vector of $\boldsymbol{b}$ with first component b_0 and the remainder those corresponding to S. When fitting the model $\boldsymbol{Y} = \boldsymbol{X}(S)\boldsymbol{b}(S) + \boldsymbol{e}'$, the least squares estimate of $\boldsymbol{b}(S)$ is $\widehat{\boldsymbol{b}}(S) = [\boldsymbol{X}(S)'\boldsymbol{X}(S)]^{-1}\boldsymbol{X}(S)'\boldsymbol{Y}$ and the minimised error sum of squares is $\mathrm{ESS}(S) = \boldsymbol{Y}'\boldsymbol{Y} - \widehat{\boldsymbol{b}}(S)'\boldsymbol{X}(S)'\boldsymbol{Y}$. We show how the Δ_k(S) can be calculated efficiently by matrix multiplications only once we have inverted $\boldsymbol{X}(S)'\boldsymbol{X}(S)$ and have $\widehat{\boldsymbol{b}}(S)$ and $\mathrm{ESS}(S)$ available.

Consider first $i \notin S$. We need to calculate $\widehat{\boldsymbol{b}}(S \cup i)$ and $\mathrm{ESS}(S \cup i)$. We may write $\boldsymbol{X}(S \cup i) = [\boldsymbol{X}(S)\boldsymbol{X}_i]$ so that

$$\begin{gathered} \boldsymbol{X}(S \cup i)'\boldsymbol{X}(S \cup i) = \begin{bmatrix} \boldsymbol{X}(S)'\boldsymbol{X}(S) & \boldsymbol{X}(S)'\boldsymbol{X}_i \\ \boldsymbol{X}_i'\boldsymbol{X}(S) & \boldsymbol{X}_i'\boldsymbol{X}_i \end{bmatrix} \\ \text{and } \boldsymbol{X}(S \cup i)'\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{X}(S)'\boldsymbol{Y} \\ \boldsymbol{X}_i'\boldsymbol{Y} \end{bmatrix}. \end{gathered} \tag{A.4}$$

Applying the formulas (A.1)–(A.3) with $\boldsymbol{A} = \boldsymbol{X}(S \cup i)'\boldsymbol{X}(S \cup i)$ and taking

$$d_{22} = d(S, i) = 1 \Big/ \left(\boldsymbol{X}_i'\boldsymbol{X}_i - \boldsymbol{X}_i'\boldsymbol{X}(S)\left(\boldsymbol{X}(S)'\boldsymbol{X}(S)\right)^{-1}\boldsymbol{X}(S)'\boldsymbol{X}_i\right) \tag{A.5}$$

eventually leads to

$$\widehat{\boldsymbol{b}}(S \cup i) = \begin{bmatrix} \boldsymbol{B}_1(S, i) \\ B_2(S, i) \end{bmatrix} = \begin{bmatrix} \widehat{\boldsymbol{b}}(S) - \left(\boldsymbol{X}(S)'\boldsymbol{X}(S)\right)^{-1}\boldsymbol{X}(S)'\boldsymbol{X}_i B_2(S, i) \\ d(S, i)\left\{\boldsymbol{X}_i'\boldsymbol{Y} - \boldsymbol{X}_i'\boldsymbol{X}(S)\widehat{\boldsymbol{b}}(S)\right\} \end{bmatrix} \tag{A.6}$$

and

$$\begin{aligned}\mathrm{ESS}(S \cup i) &= \boldsymbol{Y}'\boldsymbol{Y} - \widehat{\boldsymbol{b}}(S)'\boldsymbol{X}(S)'\boldsymbol{Y} - B_2(S,i)^2/d(S,i) \\ &= \mathrm{ESS}(S) - B_2(S,i)^2/d(S,i).\end{aligned}$$

Coming to $\Delta_i(S)$ given by (2), this implies that

$$\Delta_i(S) = B_2(S,i)^2 / \left\{ d(S,i)\sqrt{\mathrm{ESS}(S)ESS(S \cup i)} \right\} \text{ for } i \notin S.$$

These formulas are convenient: once we have $(\boldsymbol{X}(S)'\boldsymbol{X}(S))^{-1}$, $\widehat{\boldsymbol{b}}(S)$ and $\mathrm{ESS}(S)$, we can compute $d(S,i)$ and $B_2(S,i)$ by matrix multiplications only, using (A.5) and (A.6), and then $\widehat{\boldsymbol{b}}(S \cup i)$, $\mathrm{ESS}(S \cup i)$ and $\Delta_i(S)$ follow for all $i \notin S$ without the need for further matrix inversions which would have been required for each i if we used the direct formula $\widehat{\boldsymbol{b}}(S \cup i) = [\boldsymbol{X}(S \cup i)'\boldsymbol{X}(S \cup i)]^{-1}\boldsymbol{X}(S \cup i)'\boldsymbol{Y}$.

Next consider dropping an index j from S. We need to calculate $\widehat{\boldsymbol{b}}(S \setminus j)$ and $\mathrm{ESS}(S \setminus j)$ assuming again $(\boldsymbol{X}(S)'\boldsymbol{X}(S))^{-1}$, $\widehat{\boldsymbol{b}}(S)$ and $\mathrm{ESS}(S)$ known. We may now write $\boldsymbol{X}(S) = \left[\boldsymbol{X}(S \setminus j)\boldsymbol{X}_j\right]$ so that (A.4) is replaced by

$$\boldsymbol{X}(S)'\boldsymbol{X}(S) = \begin{bmatrix} \boldsymbol{X}(S \setminus j)'\boldsymbol{X}(S \setminus j) & \boldsymbol{X}(S \setminus j)'\boldsymbol{X}_j \\ \boldsymbol{X}_j'\boldsymbol{X}(S \setminus j) & \boldsymbol{X}_j'\boldsymbol{X}_j \end{bmatrix}$$

$$\text{and } \boldsymbol{X}(S)'\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{X}(S \setminus j)'\boldsymbol{Y} \\ \boldsymbol{X}_j'\boldsymbol{Y} \end{bmatrix}.$$

Applying the formulas (A.1)–(A.3) with $\boldsymbol{A} = \boldsymbol{X}(S)'\boldsymbol{X}(S)$, we now have $\boldsymbol{D} = \boldsymbol{A}^{-1} = (\boldsymbol{X}(S)'\boldsymbol{X}(S))^{-1}$ known, and therefore get $(\boldsymbol{X}(S \setminus j)'\boldsymbol{X}(S \setminus j))^{-1}$ from the last formula in (A.3) requiring only matrix multiplications. Then

$$\widehat{\boldsymbol{b}}(S \setminus j) = (\boldsymbol{X}(S \setminus j)'\boldsymbol{X}(S \setminus j))^{-1}\boldsymbol{X}(S \setminus j)'\boldsymbol{Y}$$

and $\mathrm{ESS}(S \setminus j) = \boldsymbol{Y}'\boldsymbol{Y} - \widehat{\boldsymbol{b}}(S \setminus j)'\boldsymbol{X}(S \setminus j)'\boldsymbol{Y}$ follow again requiring no matrix inversions. Then $\Delta_j(S) = [\mathrm{ESS}(S \setminus j) - \mathrm{ESS}(S)]/\sqrt{\mathrm{ESS}(S)\,\mathrm{ESS}(S \setminus j)}$.

Another useful feature of the method above is that we can compute $\boldsymbol{Y}'\boldsymbol{Y}$, the full vector $\boldsymbol{X}'\boldsymbol{Y}$ and matrix $\boldsymbol{X}'\boldsymbol{X}$ before we start with any variable selection search. All the items involved in the above calculations are sub-vectors and sub-matrices of these and therefore need not be recomputed repeatedly. In particular, each summation over observations n is done once only initially and this saves much subsequent computation time, especially for large data sets.

References

Akaike, H. (1992). Information theory and an extension of the maximum likelihood principle. *In* Kotz, S. and Johnson, N. L. (Editors) *Breakthroughs in Statistics*, volume I. Springer, New York, 610–624.

Desboulets, L. D. D. (2018). A review on variable selection in regression analysis. *Econometrics*, **6**, 45.

Dupuis, D. J. and Victoria-Feser, M.-P. (2013). Robust VIF regression with application to variable selection in large data sets. *Annals of Applied Statistics*, **7**, 319–341.

Freijeiro-González, L., Febrero-Bande, M., and González-Manteiga, W. (2022). A critical review of LASSO and its derivatives for variable selection under dependence among covariates. *International Statistical Review*, **90**, 118–145.

Grömping, U. (2007). Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician*, **61**, 139–147.

Heinze, G., Wallisch, C., and Dunkler, D. (2018). Variable selection–a review and recommendations for the practicing statistician. *Biometrical Journal*, **60**, 431–449.

Lin, D., Foster, D. P., and Ungar, L. H. (2011). VIF regression: A fast regression algorithm for large data. *Journal of the American Statistical Association*, **106**, 232–247.

Mielniczuk, J. and Teisseyre, P. (2014). Using random subspace method for prediction and variable importance assessment in linear regression. *Computational Statistics & Data Analysis*, **71**, 725–742.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley & Sons, Hoboken, NJ.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Talbot, D. and Massamba, V. K. (2019). A descriptive review of variable selection methods in four epidemiologic journals: There is still room for improvement. *European Journal of Epidemiology*, **34**, 725–730.

Wu, Y., Boos, D. D., and Stefanski, L. A. (2007). Controlling variable selection by the addition of pseudovariables. *Journal of the American Statistical Association*, **102**, 235–243.

Zhang, T. (2008). Adaptive forward-backward greedy algorithm for sparse learning with linear models. *In* Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L. (Editors) *Advances in Neural Information Processing Systems*, volume 21. 1921–1928.

Manuscript received 2021-11-04, revised 2022-02-06, accepted 2022-10-30.