Stepwise Regression under Factor Sparsity using a new Stopping Criterion

Oddgeir Samset
Department of Mathematical Sciences
The Norwegian University of Science and Technology
N-7034 Trondheim, Norway

October 29, 1998

Abstract

This paper considers variable selection in regression when both linear and higher order terms, such as cross product and quadratic terms, are taken into account. The total number of terms is then often too large to perform an exhaustive search. One approach, that enables a fairly systematic search within reasonable time, is to divide the factor space into smaller subsets, and then perform stepwise regression on each of these subsets. The subsets are here defined to include all linear and higher order terms for the different combinations of factors. A new stopping criterion in stepwise regression, based on the multiple coefficient of determination, allows all these models to be compared at the same 'level of significance'. The proposed variable selection methodology is demonstrated for both a designed experiment and an observational study.

Key words: Regression analysis; Screening; Variable selection.

1 Introduction

Variable selection in regression is widely used in the analysis of data, especially from observational studies but also from designed experiments. Such analysis are performed to identify models for prediction or explanation, or for screening the most 'important' factors. A number of methods are available, ranging from exhaustive search that examines all possible combination of terms, to stepwise methods where terms are added or removed one at a time.

For a moderate number of factors it is often possible to perform an exhaustive search in the linear terms within reasonable computing time. However, in addition to these linear terms one often wants to include higher order terms, such as cross product and quadratic terms, into the analysis. This typically increases the total number of terms extensively, and an exhaustive search might therefore not longer be feasible. One possible solution to this problem is to divide the factor space into appropriate subsets, and then use stepwise regression on each of these subsets. There are two main issues with such a procedure. The first is related to the choice of appropriate subsets. This will be discussed in Section 2. The second concerns the ability of the stopping criterion in stepwise regression to enable models from different subsets to be compared to each other. This is the topic of Section 3. In Section 4 we demonstrate the proposed variable selection procedure on data from both a designed experiment and an observational study.

2 Subsets based on Factor Sparsity

In situations where attention is given to both linear and higher order terms, such as cross product and quadratic terms, the total number of terms typically grows very large. It is thus less feasible to perform an exhaustive search through all these terms using for example *best subsets regression* due to extensive computing time. Often 'cheap' methods like *stepwise regression* are used, but these methods give however a very limited search and may result in rather poor models. Something between these two extreme strategies is requested to achieve a fairly systematic search within reasonable time. This can be obtained by dividing the factor space into appropriate subsets of a rather small number of terms, and then perform a further search using stepwise regression within each of these subsets.

At least for a designed experiment, one might argue for the *principle of factor sparsity*, which states that only a few of the factors regarded initially will contribute significantly to the variation in the response, see Box and Meyer (1986). Hence, it seems reasonable to look for subsets that contain one, two, three and so on *active* factors. Such subsets may contain terms of different types such as linear, cross product or even quadratic terms. Examples of subsets where the first two categories are included are shown below:

One active factor:	$\{x_i\}$	$i = 1, \dots, p$
Two active factors:	$\{x_i, x_j, x_i x_j\}$	$i = 1, \dots, p-1;\ j > i$
Three active factors:	$\{x_i, x_j, x_k, x_i x_j, x_i x_k, x_j x_k\}$	$i = 1, \dots, p-2;\ j > i;\ k > j$

Use of similar subsets was proposed by Box and Meyer (1993) for analysing non-replicated fractional factorial designs. Box and Meyer used a Bayesian approach to compute the marginal probabilities that the individual factors are active. They did not attempt to reduce the number of terms in a given subset and included all these into the models.

While the principle of factor sparsity might be appropriate for a designed experiment, this does not have to be the case for an observational study. There is, nevertheless, a need for dividing the total number of terms into smaller subsets in order to obtain a more thorough search for good models. Using subsets based on factor sparsity also seem reasonable in this situation. Observe that with all higher order terms removed from the subsets above we are back to an exhaustive search in the linear terms. Actually, best subsets regression in the linear terms can be viewed as a special case where one start by entering the linear terms and then specify the stopping criteria in stepwise regression to prevent any further addition and removal of terms. Hence a search using the given subsets can be seen as an extension of an exhaustive search in only the linear terms to also include higher order terms, and where stepwise regression is used to select the most significant terms in the subsets.

There are a number of pleasant advantages with this approach. We obtain a fairly systematic search through the factor space. The restrictions on the terms in a given subset tend to put less weight on less plausible models, for example models that only contain cross product terms. Stepwise regression also takes the 'level of significance' of the individual terms in the *submodels* into account, and will therefore prevent extraneous terms. This again increases the precision and prevents over-fitting. Due to the rather limited number of terms in a submodel, stepwise regression will not be too time consuming. Hence, the total computing time is only slightly larger that performing best subsets regression in the linear terms.

3 Stopping Criterion for Comparing Models

After having divided the total number of terms into appropriate subsets, we perform stepwise regression on each of these subsets. This results in a large number of models that should be compared to each other. It is therefore desirable to use a stopping criterion in stepwise regression where the 'level of significance' of the terms in different submodels are comparable. This might be a problem using the well known 'F-to-enter' and 'F-to-remove' criteria often used.

Let $Y_1, \ldots, Y_n$ be n response values and $x_1, \ldots, x_p$ a set of regression variables. Further, let the total and regression sum of squares be given in the following notation

$$\begin{aligned} SS_T &= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = SS_R + SS_E \\ SS_R &= SS_R(\beta_0, \beta_1, \beta_2, \ldots) = SS(\beta_1|\beta_0) + SS(\beta_2|\beta_1, \beta_0) + \ldots \end{aligned}$$

In standard forward regression, given that x_1 is in the model, the term x_2 with largest $SS(\beta_2|\beta_1, \beta_0)$ is included into the model if

$$F_{\text{IN}} = \frac{SS(\beta_2|\beta_1, \beta_0)/1}{SS_E(\beta_2, \beta_1, \beta_0)/(n-3)} > F_{\text{ENTER}}.$$

Similarly, let us in standard backward regression assume that x_1 and x_2 already are in the model. The term x_2 with smallest $SS(\beta_2|\beta_1, \beta_0)$ will be excluded from the model if

$$F_{\text{OUT}} = \frac{SS(\beta_2|\beta_1, \beta_0)/1}{SS_E(\beta_2, \beta_1, \beta_0)/(n-3)} < F_{\text{REMOVE}}.$$

The problem with $F_{\text{IN/OUT}}$ here is the use of MS_E as an estimator for σ^2. Hence, the 'level of significance' in each of the submodels is based on the individual estimates of σ^2. In addition, $F_{\text{IN/OUT}}$ as stopping criteria often seem to be rather unstable in the sense that they tend to include either too many or too few terms. When the model contains too few terms, MS_E often over-estimates σ^2. As a result F_{IN} tends to be too small, which again implies that F_{ENTER} must be specified small to enter 'enough' terms. Then there is a danger that also inert terms are added to the model. As a consequence MS_E often under-estimates σ^2 resulting in values of F_{IN} that are too large. This further increases the danger of adding inert terms into the model.

Assume now that σ^2 is known. It then seems natural to let the entering or removal of terms be determined by

$$\begin{aligned} G_{\text{IN/OUT}} &= \frac{SS(\beta_1|\beta_0)/1}{\sigma^2} = \frac{SS_T}{\sigma^2}\frac{SS(\beta_1|\beta_0)/1}{SS_T} = CR_1^2 \\ G_{\text{IN/OUT}} &= \frac{SS(\beta_2|\beta_1, \beta_0)/1}{\sigma^2} = \frac{SS_T}{\sigma^2}\frac{SS(\beta_2|\beta_1, \beta_0)/1}{SS_T} = C(R_2^2 - R_1^2) \\ G_{\text{IN/OUT}} &= \frac{SS(\beta_3|\beta_2, \beta_1, \beta_0)/1}{\sigma^2} = \frac{SS_T}{\sigma^2}\frac{SS(\beta_3|\beta_2, \beta_1, \beta_0)/1}{SS_T} = C(R_3^2 - R_2^2). \end{aligned}$$

Thus, by using the difference in the coefficient of multiple determination, $\Delta R_i^2 = R_i^2 - R_{i-1}^2$, we obtain the same stopping criterion as if σ^2 was known. It is straightforward to show that the order in which terms are entered or removed is exactly the same based on $F_{\text{IN/OUT}}$ or

$\Delta R^2_{\mathrm{IN/OUT}}$. Thus, both standard forward and backward regression based on $\Delta R^2_{\mathrm{IN/OUT}}$ will behave as for $F_{\mathrm{IN/OUT}}$ except for a possible different truncation point. However, due to this possible different truncation point the resulting model from stepwise regression does not have to be the same.

Since σ^2 is unknown we can not find the proportionality constant C. However, C depends only on SS_T and σ^2, and is therefore independent of which of the subsets we are examining. The terms entered or removed based on $\Delta R^2_{\mathrm{IN/OUT}}$ therefore have the same 'level of significance' based on the same value of the unknown σ^2 for all submodels. This is different for $F_{\mathrm{IN/OUT}}$, where the 'level of significance' based on the submodels MS_E, vary from submodel to submodel. The problem of choosing a reasonable value of $\Delta R^2_{\mathrm{ENTER/REMOVE}}$ is much the same as choosing $F_{\mathrm{ENTER/REMOVE}}$. However, our main intention is here to compare a large number of models and identify models that both are plausible and reasonable good according to some criteria. Even if σ^2 was known the level for entering or removing new terms would be varied to 'filter' out the most plausible models. It thus seems reasonable to start with rather small values of $\Delta R^2_{\mathrm{ENTER/REMOVE}}$, and increase these values until the number of terms in the model and the residual mean square error are appropriately small. Hopefully a few of the submodels separate out from the rest. The most plausible models found should then be examined more closely using residual analysis.

Use of the coefficient of multiple determination is a well-known tool for variable selection. If there are k terms, $\binom{k}{i}$ values of R_i^2 are computed for every number of terms i in the model, and $\max R_i^2$ for every i is plotted versus i. This plot typically shows a curve that initially increases rapidly and then remains flat. The number of terms i to include into the model is then subjectively identified to be where the 'knee' occurs. However, use of the difference in the coefficient of multiple determination as a stopping criterion in stepwise regression is to our knowledge not known.

4 Examples

4.1 Example 1: Plackett-Burman Design

Let us consider a designed experiment where the 12 run Plackett-Burman design (Plackett and Burman, 1946) is used to screen 11 factors. The objective of the experiment is to identify the active factors, i.e. those factors that are responsible for the main variation in the response. Let the response be simulated from the model $y = x_1 + x_3 + x_2x_3 + x_3x_4 + \epsilon$ with $\epsilon \sim N(0, 0.7^2)$. The design and the realisation of the response values are given in Table 1.

We first illustrate the difference between F_{IN} and ΔR^2_{IN} as stopping criteria. For the time being we therefore leave out of account the inert factors $x_5, \cdots, x_{11}$. Table 2 shows the performance of the forward selection procedure using F_{IN} as a stopping criterion when selection is among the main effects and the two-factor interactions for the four active factors, i.e. the ten regression variables $x_1, \ldots, x_4, x_1x_2, \ldots, x_3x_4$. The term x_3 is entered first into the model with a value of $F_{\mathrm{IN}} = 5.25$. This model accounts for 34.41% of the total variation, and the estimated variance is $s^2 = (1.63)^2$. With F_{ENTER} chosen to be 2.53 or less, x_2x_3, x_1 and x_3x_4, which are the rest of the terms in the true model, are included in the three next steps. However, with this 'low' value of F_{ENTER} as many as four inert terms are also included into the model before the forward selection procedure terminates. Thus, even though we here only consider the four active factors, and the correct terms enter the model in the four first steps, the use of F_{IN} as a stopping criterion either terminates at a model with one term or a model with eight terms. As a comparison Table 2 also contains the values of G_{IN} when we assume that $\sigma = 0.7$ is known. We observe that the values of G_{IN} are relatively much larger for the active than for the inert terms compared to F_{IN}. We also observe that while the values of F_{IN} increase as the first inert terms are entered into the model, the values of G_{IN} decrease. In Table 2 we have also listed the values of ΔR^2_{IN} which is proportional with G_{IN} having $C = 1.21$.

Additional 100 responses were simulated from the same model as above, and forward selection was used to select between the ten terms containing the four active factors as before. Here the four correct terms, x_1, x_3, x_2x_3 and x_3x_4, are entered into the model as the first 4 terms in 74 of the 100 simulations. They are found among the first 5, 6 and 8 terms first entered into the model in 17, 6 and 3 of the simulations, respectively. If we look closer at the 74 simulations where the four correct terms are entered first into the model, the use of ΔR^2_{IN} enables us to stop at the correct model in 73 of the simulations provided $\Delta R^2_{\mathrm{ENTER}}$ is given a reasonable value in each situation. In only one of the simulations one inert term was added before the entering of terms stopped. However, the use of F_{IN} only enables us to terminate at the true model in 33 of the 74 simulations. In 18 of the simulations one inert term was added before the entering of terms stopped. As many as 2, 4 and 5 extra terms was added for 14, 5 and 4 of the simulations, respectively. Thus, the situation presented in Table 2 seems to be fairly representative for what happens in this situation.

Table 3 gives the four 'best' models according to residual mean square error for zero to four regression variables using best subsets regression. This table shows some interesting results. First, there are three inert terms that are more correlated with the response and hence have a higher value of R^2 than the active term x_3. This might be a problems when using forward regression among all 66 regression variables. Secondly, all models listed except one contain only two-factor interactions. Such models are often regarded as less plausible. Thirdly, models with four and even three terms tend to over-fit the data. The true model is actually rated as number 217 among the models with four terms. In Table 4 we have also listed the results when forward regression is used to select between all 66 regression variables. The models found correspond with the 'best' models from best subsets regression at least for the four first regression variables. However, none of the active terms are found. Stepwise regression would give no improvement to forward regression here since the term last entered always is the candidate to be removed. Hence, both best subsets regression and stepwise regression on all 66 regression variables seem to be inappropriate in this situation, and fail to identify the active factors.

Table 5 and Table 6 give the results of forward submodel regression based on F_{IN} and ΔR^2_{IN}, respectively. The chosen values for F_{ENTER} and $\Delta R^2_{\mathrm{ENTER}}$ are here 2.50 and 14.30, respectively. Both these values are based on the information from Table 2. If they are increased any higher the correct model will not be found, and gives therefore the 'optimal' tuning of the stopping criteria. For both stopping criteria most of the models listed now contain both main effects and two-factor interactions. When using F_{IN} the model containing the four active factors are rated as number five. However, this model also contains the four extra inert terms as displayed in Table 2. If we increase F_{ENTER}, as might seem naturally here due to the large number of terms for four factors, the true model will not be found. When using $\Delta R^2_{\mathrm{ENTER}}$ equal to 14.3 the correct model is rated as number two among the models containing four potentially active factors. The model rated as number one, however, contains only two-factor interactions. When specifying $\Delta R^2_{\mathrm{ENTER}}$ equal to 10.0 the true model will be rated as number five. If $\Delta R^2_{\mathrm{ENTER}}$ is as low as 5.00 the true model is rated as number 30.

However, then there are a large number of models with relatively many terms giving about the same fit to the data. The general impression in this example is that the models found using F_{IN} contain a rather large number of terms compared to the models found by using ΔR^2_{IN}.

4.2 Example 2: The Detroit Data

The data we use here are from an observational study on the homicide rate in Detroit 1961 to 1973. The data were collected and used by Fisher (1976), and are given in Appendix A in Gunst and Mason (1980). The data are also available from StatLib. It contains 13 observation of 3 different responses and 11 regression variables, see Table 7. The response variable used here is the number of homicides per 100,000 of population (y_1). The main objective is to identify models to be used for prediction.

Miller (1990) used these data to compare several variable selection techniques, and pointed out that there exist a model with three linear terms found by best subsets regression that seems to give much better fit that the models found by other variable selection methods. While Miller only consider the linear terms we also include the 55 cross product terms. Note that the 11 linear terms are here first standardised to have zero mean and sample standard deviation equal to one.

Table 8 shows the results from best subsets regression. The five models with least residual mean square error for zero to three terms are given. We observe that by introducing cross product terms into the analysis there are actually two models with three terms giving a rather good fit to the data having values of s about 0.86. From Table 9 we see that the best model with tree terms, containing x_6, x_4 and x_{10}, only have $s = 1.53$ when forward regression is applied on all 66 regression variables. By using stepwise regression with $F_{\text{ENTER}} = F_{\text{REMOVE}}$ equal to 5.35, the term x_6 will be removed in step 6. Hence, we can find the two-term model (x_6, x_4) and the four-term model (x_4, x_{10}, x_1x_9, x_3x_6). By using $\Delta R^2_{\text{ENTER}} = \Delta R^2_{\text{REMOVE}}$ equal to 0.35 we can also find the three-term model (x_4, x_{10}, x_1x_9) which is rated as number four among among the three-terms models from Table 8. The three models rated first can not be found.

Table 10 gives the results for forward submodel regression using $F_{\text{ENTER}} = 4.0$. The two best models with three terms from best subsets regression are now found. Many other of the models from best subsets regression, or models that are differing in only one term, are also found. Thus, the proposed methodology have identified a number of alternative models that should be examined more closely.

In this situation there are rather small differences in which models that are found using either F_{IN} or ΔR^2_{IN}. Note however the difference in the behaviour for these two stopping criteria from Table 9. While the values of F_{IN} are fluctuating between high and low values, ΔR^2_{IN} seem to be much more stable decreasing as terms are added into the model. If the response instead are chosen to the death rate in accidents per 100,000 population (y_2), the difference between these two criteria will be larger. Again, as for the designed experiment in the previous sections, F_{IN} then tend to include more terms than ΔR^2_{IN}.

References

Box, G. E. P. and Meyer, R. D. (1986). An Analysis for Unreplicated Fractional Factorials, *Technometrics* **28**, 11–18.

Box, G. E. P. and Meyer, R. D. (1993). Finding the Active Factors in Fractionated Screening Experiments, *Journal of Quality Technology* **25**, 94–105.

Fisher, J. C. (1976). Homicide in Detroit: the role of firearms, *Criminology* **14**, 387–400.

Gunst, R. F. and Mason, R. L. (1980). *Regression Analysis and its Application*, Marcel Dekker, New York.

Miller, A. J. (1990). *Subset Selection in Regression*, Chapman and Hall, New York.

Plackett, R. L. and Burman, J. P. (1946). The Design of Optimum Multifactorial Experiments, *Biometrica* **33**, 305–325.

Table 1: The 12 run Plackett-Burman design and response used in Example 1.

run	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	y
1	+	-	+	-	-	-	+	+	+	-	+	-0.416
2	+	+	-	+	-	-	-	+	+	+	-	-1.909
3	-	+	+	-	+	-	-	-	+	+	+	0.025
4	+	-	+	+	-	+	-	-	-	+	+	1.563
5	+	+	-	+	+	-	+	-	-	-	+	-2.378
6	+	+	+	-	+	+	-	+	-	-	-	3.316
7	-	+	+	+	-	+	+	-	+	-	-	1.297
8	-	-	+	+	+	-	+	+	-	+	-	-0.348
9	-	-	-	+	+	+	-	+	+	-	+	-3.053
10	+	-	-	-	+	+	+	-	+	+	-	1.933
11	-	+	-	-	-	+	+	+	-	+	+	-1.722
12	-	-	-	-	-	-	-	-	-	-	-	-0.369

Table 2: Behaviour of the forward procedure in Example 1. Selection is among the ten regression variables containing the four active factors.

Step	Terms entered	R^2	s	F_{IN}	G_{IN}	ΔR^2_{IN}
1	x_3	34.41	1.63	5.25	28.45	34.41
2	x_2x_3	48.78	1.52	2.53	11.88	14.37
3	x_1	67.79	1.28	4.72	15.72	19.01
4	x_3x_4	92.10	0.68	21.53	20.10	24.30
5	x_2	94.50	0.61	2.62	1.99	2.40
6	x_4	96.81	0.51	3.63	1.91	2.31
7	x_1x_2	98.63	0.37	5.32	1.50	1.82
8	x_2x_4	99.63	0.22	8.11	0.83	1.00
9	x_1x_4	99.70	0.25	0.48	0.06	0.07
10	x_1x_3	96.73	0.33	0.10	0.02	0.03

Table 3: Best subsets regression in Example 1.

R^2	R^2_{adj}	s	Terms
0.00	0.00	1.92	None
57.42	53.16	1.31	x_7x_9
43.31	37.64	1.52	x_4x_5
42.06	36.27	1.53	x_1x_6
34.41	27.85	1.63	x_3
82.59	78.72	0.89	x_7x_9 x_8x_9
75.92	70.56	1.04	x_4x_5 x_7x_9
75.06	69.51	1.06	x_1x_6 x_7x_9
74.42	68.73	1.07	x_7x_9 x_8x_{11}
98.03	97.29	0.32	x_7x_9 x_8x_9 x_8x_{11}
91.50	88.32	0.66	x_1x_3 x_2x_4 x_7x_9
89.89	86.10	0.72	x_4x_5 x_7x_9 x_8x_9
89.60	85.70	0.73	x_1x_6 x_1x_9 x_3x_{10}
99.27	98.85	0.21	x_1x_{11} x_7x_9 x_8x_9 x_8x_{11}
99.25	98.82	0.21	x_1x_6 x_1x_9 x_1x_{11} x_3x_{10}
99.17	98.70	0.22	x_7x_9 x_8x_9 x_8x_{10} x_8x_{11}
98.97	98.38	0.24	x_3x_9 x_7x_9 x_8x_9 x_8x_{11}

Table 4: Forward regression in Example 1. Selection is among all 66 regression variables.

Step	Terms entered	R^2	s	F_{IN}
1	x_7x_9	57.42	1.31	13.48
2	x_8x_9	82.59	0.89	13.01
3	x_8x_{11}	98.03	0.32	62.63
4	x_1x_{11}	99.27	0.21	11.80
5	x_7x_{11}	99.81	0.11	17.14
6	x_1x_3	99.95	0.06	13.32
7	x_2	99.99	0.03	18.47
8	x_5x_7	100.00	0.01	18.84

Table 5: Forward submodel regression in Example 1. $F_{\mathrm{ENTER}} = 2.50$.

R^2	R^2_{adj}	s	Terms
34.41	27.85	1.63	x_3
20.16	12.18	1.80	x_{11}
57.42	53.16	1.31	x_7x_9
57.73	48.34	1.38	x_6 x_1x_6
54.57	44.48	1.43	x_3 x_{11}
50.08	38.99	1.50	x_3 x_6
88.13	83.68	0.78	x_7 x_7x_9 x_8x_9
88.38	81.75	0.82	x_3 x_2x_3 x_2x_8 x_3x_8
86.92	79.45	0.87	x_3 x_3x_7 x_3x_9 x_7x_9
86.90	75.99	0.94	x_3 x_6 x_{11} x_3x_{11} x_6x_{11}
100.00	100.00	0.01	x_1 x_2 x_9 x_{11} x_1x_2 x_1x_9 x_1x_{11} x_2x_{11} x_9x_{11}
99.94	99.68	0.11	x_2 x_5 x_8 x_2x_5 x_2x_8 x_2x_{10} x_5x_8 x_5x_{10} x_8x_{10}
99.93	99.59	0.12	x_3 x_8 x_{10} x_2x_3 x_2x_8 x_2x_{10} x_3x_8 x_3x_{10} x_8x_{10}
99.83	99.04	0.19	x_1 x_6 x_8 x_1x_6 x_1x_8 x_1x_{10} x_6x_8 x_8x_{10}
99.63	98.65	0.22	x_1 x_2 x_3 x_4 x_1x_2 x_2x_3 x_2x_4 x_3x_4

Table 6: Forward submodel regression in Example 1. $\Delta R^2_{\mathrm{ENTER}} = 14.3$.

R^2	R^2_{adj}	s	Terms
34.41	27.85	1.63	x_3
20.16	12.18	1.80	x_{11}
15.67	7.24	1.85	x_6
57.42	53.16	1.31	x_7x_9
57.73	48.34	1.38	x_6 x_1x_6
54.57	44.48	1.43	x_3 x_{11}
50.08	38.99	1.50	x_3 x_6
82.59	78.72	0.89	x_7x_8 x_8x_9
82.08	75.36	0.95	x_3 x_2x_3 x_2x_8
81.73	74.89	0.96	x_3 x_{11} x_3x_7
76.74	68.02	1.09	x_6 x_1x_6 x_6x_{11}
98.03	97.29	0.32	x_7x_8 x_8x_9 x_8x_{11}
92.10	87.58	0.68	x_1 x_3 x_2x_3 x_3x_4
90.86	85.63	0.73	x_{11} x_2x_{10} x_8x_{10} $x_{10}x_{11}$
82.59	78.72	0.89	x_7x_8 x_8x_9

Table 7: Description of the Detroit data used in Example 2.

Variables	Comment	
x_1	FTP	Full-time police per 100,000 population
x_2	UEMP	% unemployed in the population
x_3	MAN	Number of manufacturing workers in thousands
x_4	LIC	Number of handgun licences per 100,000 population
x_5	GR	Number of handgun registrations per 100,000 population
x_6	CLEAR	% homicides cleared by arrests
x_7	WM	Number of white males in the population
x_8	NMAN	Number of non-manufacturing workers in thousands
x_9	GOV	Number of government workers in thousands
x_{10}	HE	Average hourly earnings
x_{11}	WE	Average weekly earnings
y_1	HOM	Number of homicides per 100,000 of population
y_2	ACC	Death rate in accidents per 100,000 population
y_3	ASR	Number of assaults per 100,000 population

Table 8: Best subsets regression in Example 2.

R^2	R^2_{adj}	s	Terms
0.00	0.00	16.39	None
93.79	93.23	4.26	x_6
92.94	92.30	4.55	x_1
91.79	91.04	4.90	x_9
91.38	90.60	5.02	x_8
90.74	89.90	5.21	x_7
98.95	98.74	1.84	x_4 x_6
98.74	98.49	2.01	x_7 x_7x_9
98.61	98.33	2.12	x_2 x_7
98.60	98.32	2.12	x_9 x_7x_8
98.42	98.10	2.26	x_7 x_4x_7
99.79	99.73	0.86	x_4 x_9 x_9x_{10}
99.79	99.72	0.87	x_2 x_4 x_{11}
99.61	99.48	1.18	x_4 x_7 x_7x_9
99.61	99.48	1.18	x_4 x_{10} x_1x_9
99.58	99.44	1.23	x_9 x_4x_7 x_6x_{10}

Table 9: Behaviour of the forward procedure in Example 2. Selection is among all 66 regression variables.

Step	Terms entered	R^2	s	F_{IN}	ΔR^2_{IN}
1	x_6	93.79	4.26	166.18	93.79
2	x_4	98.95	1.84	49.13	5.16
3	x_{10}	99.34	1.53	5.37	0.39
4	x_1x_9	99.80	0.89	18.71	0.46
5	x_3x_6	99.93	0.58	11.88	0.12
6	x_7x_8	99.95	0.52	2.86	0.02
7	x_5x_{11}	99.97	0.46	2.44	0.02
8	x_1x_5	100.00	0.11	79.0	0.03
9	x_5x_6	100.00	0.03	66.95	0.00
10	x_2x_3	100.00	0.01	50.86	0.00

Table 10: Forward submodel regression in Example 2. $F_{\mathrm{ENTER}} = 4.0$.

R^2	R^2_{adj}	s	Terms
0.00	0.00	16.39	None
93.79	93.23	4.26	x_6
92.94	92.30	4.55	x_1
91.79	91.04	4.90	x_9
91.38	90.60	5.02	x_8
90.74	89.90	5.21	x_7
99.44	99.26	1.41	x_3 x_7 x_3x_7
99.12	98.83	1.77	x_7 x_9 x_7x_9
99.08	98.78	1.81	x_9 x_{10} x_9x_{10}
98.95	98.74	1.84	x_4 x_6
98.61	98.33	2.12	x_2 x_7
99.79	99.73	0.86	x_4 x_9 x_9x_{10}
99.79	99.72	0.87	x_2 x_4 x_{11}
99.67	99.51	1.15	x_4 x_7 x_9 x_7x_9
99.66	99.49	1.17	x_5 x_9 x_5x_{10} x_9x_{10}
99.60	99.40	1.27	x_4 x_6 x_{10} x_6x_{10}