ANOTHER LOOK AT HOCKING'S GAS MILEAGE DATA

Harold V. Henderson
Biometrics Unit, Cornell University, Ithaca, N. Y. 14853

BU-619-M July 1977

Abstract

Hocking [1976] in his excellent review uses this data set to illustrate some techniques for analysis and selection of variables. We inject some subject matter considerations to lead to a more satisfactory model for prediction.

1. INTRODUCTION

The data is extracted from Motor Trend [1974] in an attempt to predict gasoline mileage (MPG) for 1973-74 automobiles. Ten variables measuring various aspects of the automobile design and performance were recorded on 32 automobiles:

Number	Variable	SAS Name
1	Engine Shape [Straight (1), V (0)]	Engshape
2	Number of cylinders	Cylinder
3	Transmission type [Manual (1), Auto (0)]	Gears
4	Number of transmission speeds	Nogears
5	Engine size	Engsize
6	Horsepower	Hp
7	Number of carburetor barrels	Carbs
8	Final drive ratio	Dratio
9	Weight	Weight
10	Quarter mile time	Qmtime

Hocking's [1976] analysis suggests that the subset (3,9,10) may be best for prediction. This model is difficult to interpret and the absence of variable 5, Engine size, is surprising. The purpose of this study is to inject some subject matter considerations (and even let the data lead) to obtain a better model for prediction and thereby answer questions posed by Hocking [1976].

2. VARIABLE ANALYSIS

Other possible factors that might influence, or could be used to predict, gas mileage in addition to the 10 variables considered could include:

a) Further aspects of load; e.g., power accessories like air conditioning, pollution devices.

b) Class of automobile; e.g., sport, luxury, compact, sub-compact, country of manufacture.

c) Further aspects of performance; e.g., power/weight ratio, maximum speed, carburation, engine tuning.

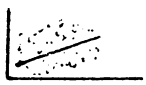
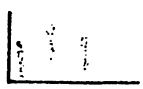
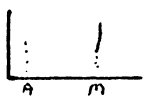
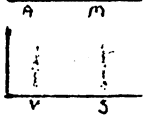
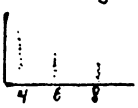
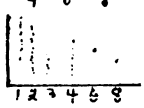
The power/weight ratio was added to the data set as variable 12 and calculated as Hp per wt. = Hp/weight.

This particular sample of 32 automobiles is hardly a "representative" sample with a bias to exotic, non-U.S., automobiles with 7 Mercedes, a Porsche, a Ferrari and a Maserati. We note also that the Mercedes 240D has a diesel engine and the 2 Mazdas with rotary engines are coded as V6 engines in Hocking's analysis. So we may not expect a very useful or portable predictor model to emerge. (We have not corrected these to enable direct comparisons with Hocking's analyses.)

Plotting is a much neglected technique in variable analysis and subset selection. Plots, using SAS76, of MPG against the 10 variables in Hocking [1976] are sketched in Table 1. These plots indicate possible functional relationships and provide visual information on the need and type of transformation required, if any. Particular note is made of differences between different models from the same

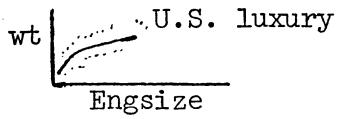
manufacturer which differ in only a few of the variables. We also look at each variable to check whether its range covers the range of interest. This is summarized in Table 1.

Table 1: Variable Summary

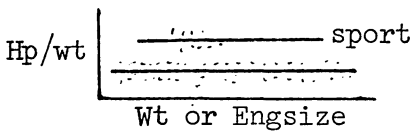
Variable*	Nature	Plot (sketch)	Relationship	Scatter	Potential
Weight	Cont.		Near straight line	Tight	Useful
Engine size	Cont.		Curvilinear	Tight	Useful
Hp	Cont.		Curvilinear	Less tight	Not as useful as Engsize; sport, luxury interchanged
Qmtime	Cont.		Straight line	Wide	Not too useful, outlier
Dratio	Cont.		Straight line	Wide	Not very useful
Number of gears	Discrete		Possibly curvilinear	Wide	Not useful in its own regard but "standin" for auto class
Transmission type	Discrete		Straight line	Wide	Not useful
Engine shape	Discrete		Straight line	Wide	Not useful
Cylinders	Discrete		Straight line	Wide	Not useful
No. of carburetors	Discrete		Curvilinear	Wide	Not too useful; "standin" for auto class
Hp/wt	Cont.		Curvilinear	Wide	Lotus outlier, not too useful

*The range of all variables was "good".

Multicollinearity of the independent variables is indicated by the high single correlation coefficients in Table 2 of Hocking [1976], which indicate straight line relationships existing among some of the variables. Evidence for curvilinear relationships are gained by plotting pairs of independent variables. For example, the plot of weight against engine size shows a strong curvilinear relationship. Hp/wt against engine size or weight shows two straight line relationships for sport and non-sport automobiles. We are getting an indication that a "class of automobile" variable should have been included.



Weight vs. Engsize

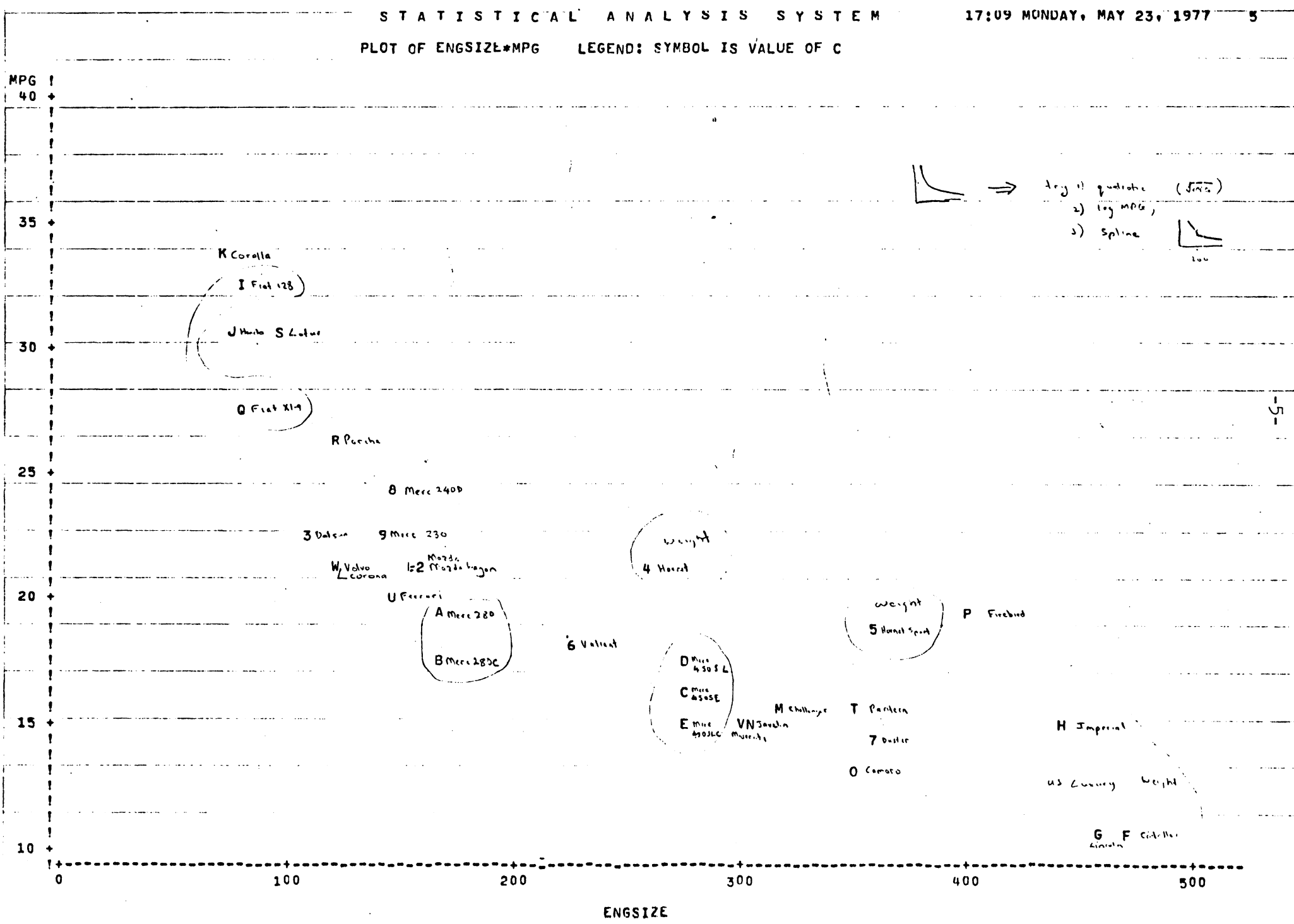


Hp/wt vs. Engsize or Weight

Hocking [1976] points out that from eigenvalue analysis of X'X there is a strong linear relationship between Engsize, Carburetors, Weight (5,7,9) and possibly between No. of Cylinder, Hp, Qmtime (2,6,10). This together with our discussion of the usefulness of each variable would suggest dropping No. of Carburetors (7) and No. of Cylinders (2).

The relationships between MPG and the independent variables show straight line relationships for most of the variables. Subject matter considerations suggest engine size to be important. The curvilinear relationship (Fig. 1) with MPG suggests using a transformation. Again plotting was the tool. Log (MPG) was better than $(MPG)^{\frac{1}{2}}$ but $(MPG)^{\frac{1}{3}}$ was best, indicating a cubic polynomial in engine size would be worth trying. (Transforming Engsize rather than MPG because of its straight line relationships with other independent variables, and stopping at cubic because the quartic term was not significant.)

Figure 1


STATISTICAL ANALYSIS SYSTEM 17:09 MONDAY, MAY 23, 1977 5
PLOT OF ENGSIZE*MPG LEGEND: SYMBOL IS VALUE OF C
MPG
40
35
30
25
20
15
10
0
100
200
300
400
500
ENGSIZE
K Corolla
I Fiat 128
J Honda S Colt
Q Fiat X1-9
R Porsche
8 Merc 240D
3 Datsun
9 Merc 230
W Volvo
Corona
1 2 Mazda
Mazda Wagon
U Ferrari
A Merc 280
B Merc 280C
6 Valiant
weight
4 Hornet
D Merc 450SL
C Merc 450SE
E Merc 450SLC
VN Javelin
M Challenger
T Pantera
7 Duster
O Camaro
weight
5 Hornet Sport
P Firebird
H Imperial
us Luxury Weight
G F Cadillac
Lincoln
log ⇒ 1) quadratic (√ENGSIZE)
2) log MPG,
3) Spline
Too

Thus, examination of the plots and subject matter considerations suggest discarding 7 variables:

3 Transmission type

1 Engine shape

2 No. of Cylinders

7 No. of Carburetors

8 Dratio

12 Hp/wt

10 Qmtime

and adding 2 variables:

11 (Engine size)^2

13 (Engine size)^3

leaving the 6 variables:

5 Engine size

11 (Engine size)^2

13 (Engine size)^3

9 Weight

6 Hp

4 No. of Gears

as the variables that "look good".

3. SUBSET SELECTION

We have reduced the contending variables for a prediction equation by evaluating the variables using plots and subject matter considerations. These are all pre-statistical and pre-regression/subset selection package considerations. It is now a simple matter to choose the "best" predictor subset from these six variables.

We will, however, proceed without the benefit of the reduction (but not the addition) in the data set and run all 13 variables in various regression/subset selection packages, as this is unfortunately closer to the way users proceed.

In the absence of prior knowledge of σ^2 we obtain an estimate using the "Nearest Neighbour Technique" (NNT) in order to see how well the predictor equation is doing. The nearest neighbour classes are indicated on the data printout and give an estimate of σ^2 as $S^2 = 4.46$ with 17 degrees of freedom. This is quite high and results because of the variation between automobiles from even the same manufacturer differing in only a few (and often only one) variable and yet differing by as much as 5 mpg. One has only to scan the plots noting, for example, the Mercedes 450SE (Sedan), 450SL (Roadster) and 450SLC (Coupe) with 16.4, 17.3 and 15.2 mpg, yet differing only in weight and Qmtime. Similarly, the Fiat 128 records 32.4 mpg while the sport coupe X1-9 records 27.3 mpg.

All data are analysed in standard form, so that X'X is the correlation matrix, to enable direct comparisons with Hocking's [1976] analyses. The estimate of σ^2 for the scaled data is $S_S^2 = S^2/1126.047 = 3.96^{-3}$. [The intercept is forced into the equation when the variables are standardised so we must adjust the degrees of freedom from the output (this would have been unncessary if models with intercept had been run, with SS(Intercept) = 0) by adding 1 to regression d.f. and subtracting 1 from error d.f.]

Running the "full model", i.e., with all 13 independent standardised variables gives $RMS_{14} = \hat{\sigma}^2 = 3.01 \times 10^{-3} < S_S^2 = 3.96 \times 10^{-3}$, the NNT estimate. So we are well in the ball park.

SAS76 Stepwise procedure with Forward Selection, Stepwise, Backward Elimination and Maximum R^2 were run. The results are summarised in Table 2.

Table 2: Comparison of Techniques in SAS76 Stepwise Procedure on Standardized Variables 1-13

	Forward Selection (FS)				Backward Elimination (BE)				Maximum R^2 (Max R^2)			
p	Variables[1]	Sig[2]	$RMS_p \times 10^3$	R^2	Variables	Sig	$RMS_p \times 10^3$	R^2	Variables	Sig	$RMS_p \times 10^3$	R^2
2	Weight	**	8.24	.75	Engsize	**	9.35	.72	Weight	**	8.24	.75
3	Cylinder	**	5.85	.83	$(Engsize)^2$	**	7.12	.79	Cylinder -Cylinder	**	5.85	.83
4	Hp/wt	NS	5.60	.84	$(Engsize)^3$	**	4.38	.88	Qmtime, $(Engsize)^3$ -Weight, -Qmtime	*	5.34	.85
5	$(Engsize)^3$	NS	5.05	.86	Hp	*	3.64	.90	Engsize, $(Engsize)^2$, Hp	*	3.64	.90
6	Engsize	NS	4.97	.87	Hp/wt	NS	3.42	.91	No. of gears	**	2.88	.93
7	$(Engsize)^2$	**	3.83	.90	Cylinder	NS	3.42	.91	Dratio	NS	2.81	.93
8	Hp	NS	3.52	.92	Carbs	NS	3.16	.92	Qmtime -Qmtime	NS	2.80	.93
9	No. of gears		3.01	.93	Weight		3.04	.93	Carbs, Cylinder		2.77	.93
10	Carbs		2.94	.94	Gears		2.94	.94	Qmtime		2.84	.94
11	Dratio		2.91	.94	Qmtime		2.82	.94				
12	Qmtime		2.96	.94	Engshape		2.82	.94				
13	Engshape		2.98	.94	No. of gears		2.92	.94				
14	Gears		3.01	.95	Dratio		3.01	.95				

[1] The column headed 'Variables' indicates the variables added or deleted (-) as p is increased.

[2] NS, *, ** indicate the significance level of the last variable added >5%, 5%, or 1%.

Notable features include the wide difference between FS and BE. FS and stepwise, which terminated at step 2 (p = 3), remained locked on weight and cylinder which are the best subsets for p = 2 and 3 . FS shows a "jolt" at p = 7 when $(\text{Engsize})^2$ is added. BE is very similar to Maximum R^2, giving the same subset for p = 5 with RMS = 3.64×10^{-3} . Max R^2 makes a significant leap for p = 6, giving RMS = 2.88×10^{-3} . Max R^2 picks the best subsets for p = 2, 3, is partially locked in for p = 4, but recovers fully by p = 5 .

Summarizing the discussion, we tabulate in Table 3 the best subset of each size found in these analyses together with their C_p value. Recall a subset with $C_p \leq p$ is a candidate for prediction.

Table 3

p[1]	Variables (Standardized)	$RMS_p \times 10^3$	C_p[2]
2	Weight	8.24	54.4
3	Weight, cylinder	5.85	30.6
4	Engsize, $(\text{Engsize})^2$, $(\text{Engsize})^3$	4.38	16.8
5	" " " , Hp	3.64	10.8
6	" " " " , No gears	2.88	4.9
7	" " " " " , Dratio	2.81	5.4

[1] p is the number of variables plus one for the implicit intercept.

[2] $C_p = RSS_p/\hat{\sigma}^2_{full} + 2p - n = RSS_p/3.01 \times 10^{-3} + 2p - 32$.

Note these C_p values are not comparable with Hocking's, which use $\hat{\sigma}^2_{full} = 6.24 \times 10^{-3}$

Since prediction is the name of the game, a criterion for the best predictor subset is the subset with minimum PRESS. The suggestion is to evaluate PRESS for all possible subsets. This would be $2^{13} = 8192$ in this case, which would be a

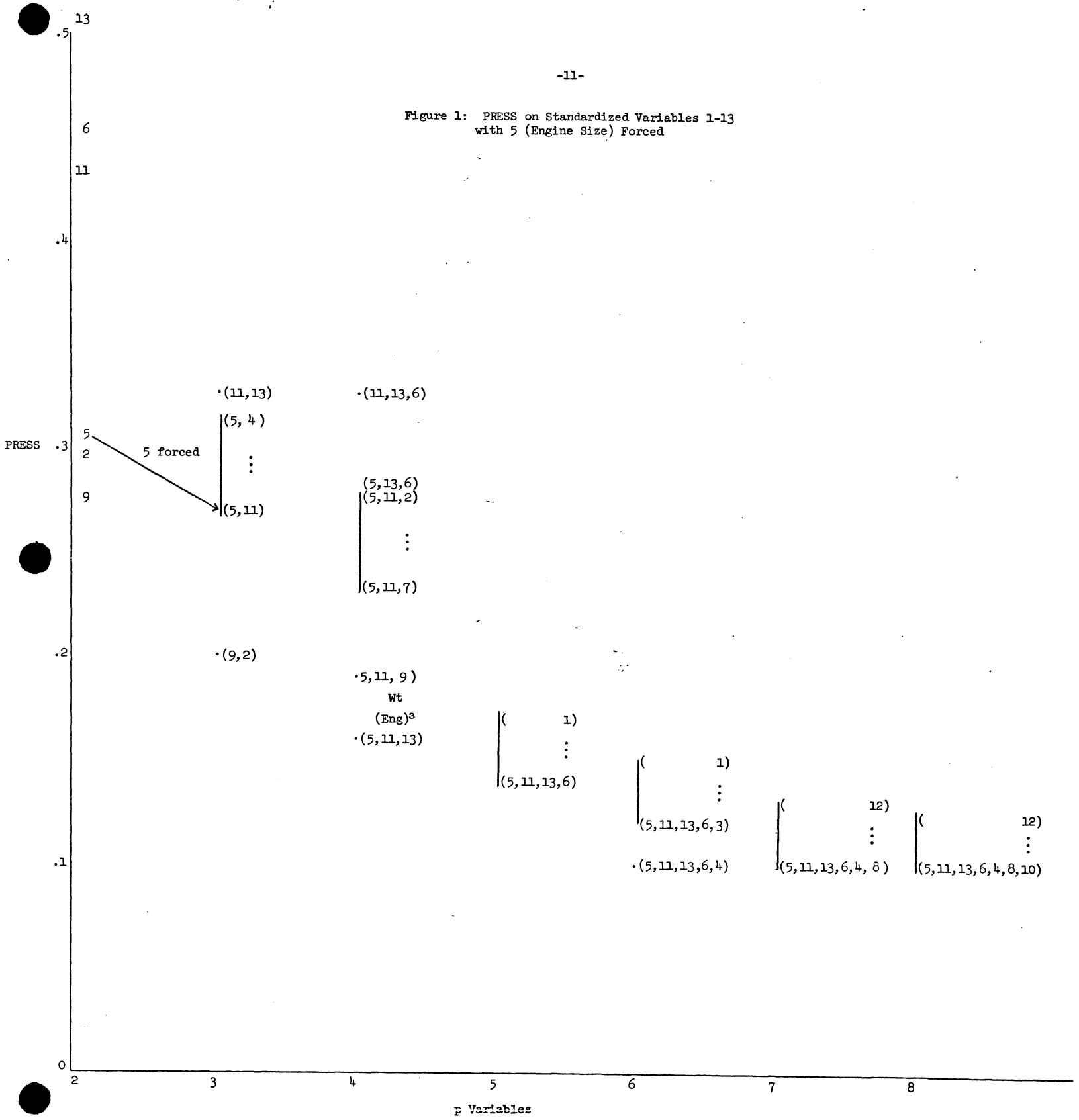
moderate amount of computing. Allen has written a stepwise procedure to be used when the number of variables is large. Using this stepwise procedure on the 13 standardised variables does not do well. Suggestions are made in Appendix 1 on modifications to improve the performance. However by forcing variable 5, Engine size, to be in the prediction equation we achieve what must be the near best subsets for $p \geq 3$, given in Table 4.

Table 4: PRESS on Standardised Variables 1-13 with Engine Size Forced

Allen's PRESS Program, Engsize Forced							
p	Variables			Min $PRESS_p$ Found	C_p	$RMS_p \times 10^3$	Best p-set Found also by
2	Engsize	5	**	.304	65.5	9.35	BE
3	$(Engsize)^2$	11	**	.252	43.1	7.12	BE
4	$(Engsize)^3$	13	**	.146	16.8	4.38	BE
5	Hp	6	*	.131	10.8	3.64	BE=Max R^2
6	No. of gears	4	**	.109	4.9	2.88	Max R^2
7	Dratio	8	NS	.106	5.4	2.81	Max R^2
8	Qmtime	10	NS	.106	6.4	2.80	Max R^2

The graph (Fig. 1) of $PRESS_p$ for the subsets considered shows rapid decrease until $p = 6$ then levelling off. This suggests the same subset (5,11,13,6,4) as Max R^2 does. So we have some degree of confidence in it being a good condidate for prediction. This is a cubic polynomial in engine size with Hp and Number of gears. The number of gears variable might be regarded as a "stand-in" for automobile class with S = sport, 4 = compact, mainly import, 5 = luxury, family. It might be expected that an automobile class variable would do better.

Figure 1: PRESS on Standardized Variables 1-13
with 5 (Engine Size) Forced



The residual plot for this predictor equation looks good, having a good scatter with no apparent pattern.

4. CONCLUSION

Hocking [1976] posed the question of why the ridge regression analysis "judges as unreliable two of the [his] three 'essential' variables". We have seen that Hocking's analysis does not contain all relevant variables (and neither does ours), but it appears that taking the curvilinear relationship of MPG with Engine size leads to much improved prediction subset. We note that a cubic polynomial on Engine size alone does better than Hocking's best subset (3,9,10) with $RMS_4 \times 10^3$ of 4.38 and 5.37 respectively, and that neither of these is less than the nearest neighbour estimate of $\sigma^2 \times 10^3$, 3.96.

The PRESS and Max R^2 analyses for $p = 6$ coincide and this subset (5,11,13,6,4) appears to be the best for prediction based on:

1 PRESS levelling off,

2 Best $p = 6$ subset for Max R^2,

3 RMS levelling off,

4 Significance of β's at 1% level,

5 Residual plot,

6 $C_6 = 4.9 \leq 6 \Rightarrow$ prediction candidate,

7 Subject matter considerations,

8 $RMS_6 \times 10^3 = 2.88 < 3.96 = S_S^2 \times 10^3$ from nearest neighbour technique.

The final validation would be an evaluation based on another data set.

Packages and selection routines are no substitute for a combination of subject matter considerations and common sense. But a combination of these can be a very powerful technique.

ACKNOWLEDGMENTS

I would like to thank Dr. F. B. Cady for presenting valuable course material on variable selection which motivated and is implicit in this study, and Dr. O. P. Hackney for the data set. Computing money was provided by Hatch Fund 402, Biometrics Unit, Cornell University, while the author was supported by a New Zealand National Advisory Council Post-Graduate Research Fellowship.

REFERENCES

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

APPENDIX

Improvements to the PRESS and C_p Stepwise Routine

Presently the stepwise procedure is weighted in favour of adding rather than deleting variables at each stage. The subset of variables with minimum PRESS (C_p) is chosen from a set where the subsets with variables being added have two more variables than those where a variable is deleted. This can be improved by comparing the subsets where a variable is being deleted with its own, rather than having it outclassed. This simply requires keeping account of the subset with minimum PRESS (C_p) for each size subset.

The program at present performed badly on the gas mileage data set. p-subsets with smaller PRESS than those p-subsets worked were calculated on delete mode but not worked because they were compared only with subsets with two more variables. These are the subsets in bold on the graph (Fig. 2) of PRESS on the 13 standardized variables. Figure 2 is to be compared with Figure 1 in the paper, where PRESS values down to .106 were achieved.

The modifications are presented in Figures 3 and 4 as a Structured Diagram.

I also had problems with C_p output. The C_p values computed were very similar to $PRESS_p$; RESS (Residual Sum of Squares) is calculated correctly but $C_p = RESS_p/\hat{\sigma}^2_{full} + 2p - n$ is not (e.g., for variable 9, Weight)

$$RESS = .247167, \quad \sigma^2_{full} = \frac{.05437375}{18} .$$

$X_2 = .247167/\sigma^2_{full} + 4 - 32 \neq .253$ as given in the output.

Figure 2: PRESS on Standardised Variables 1-13

Note: The program as it stands does not take these backward steps (in boldface type) although it does calculate PRESS for these subsets!

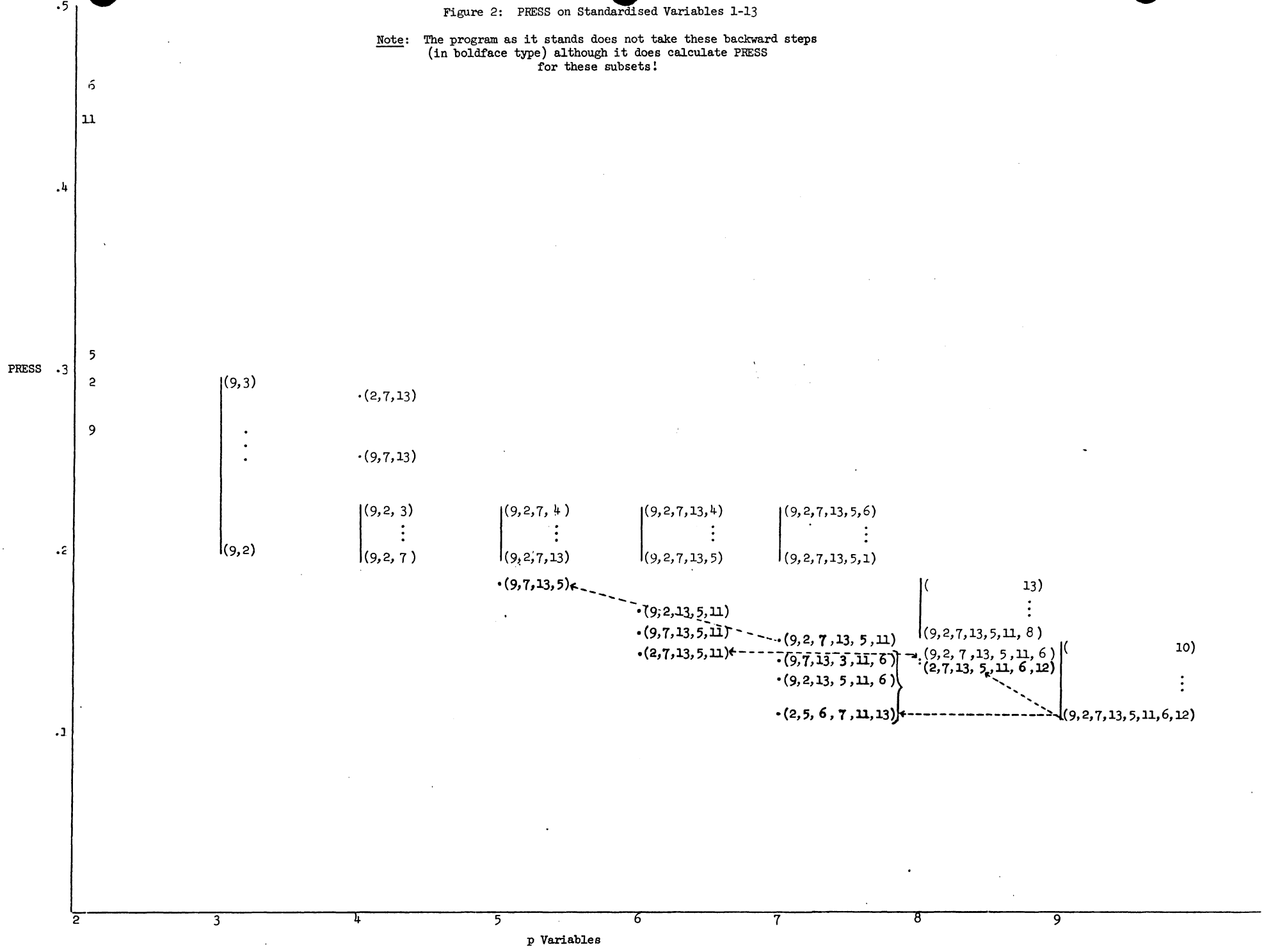


Figure 3: Structured Diagram for Modification to PRESS and C_p Routine

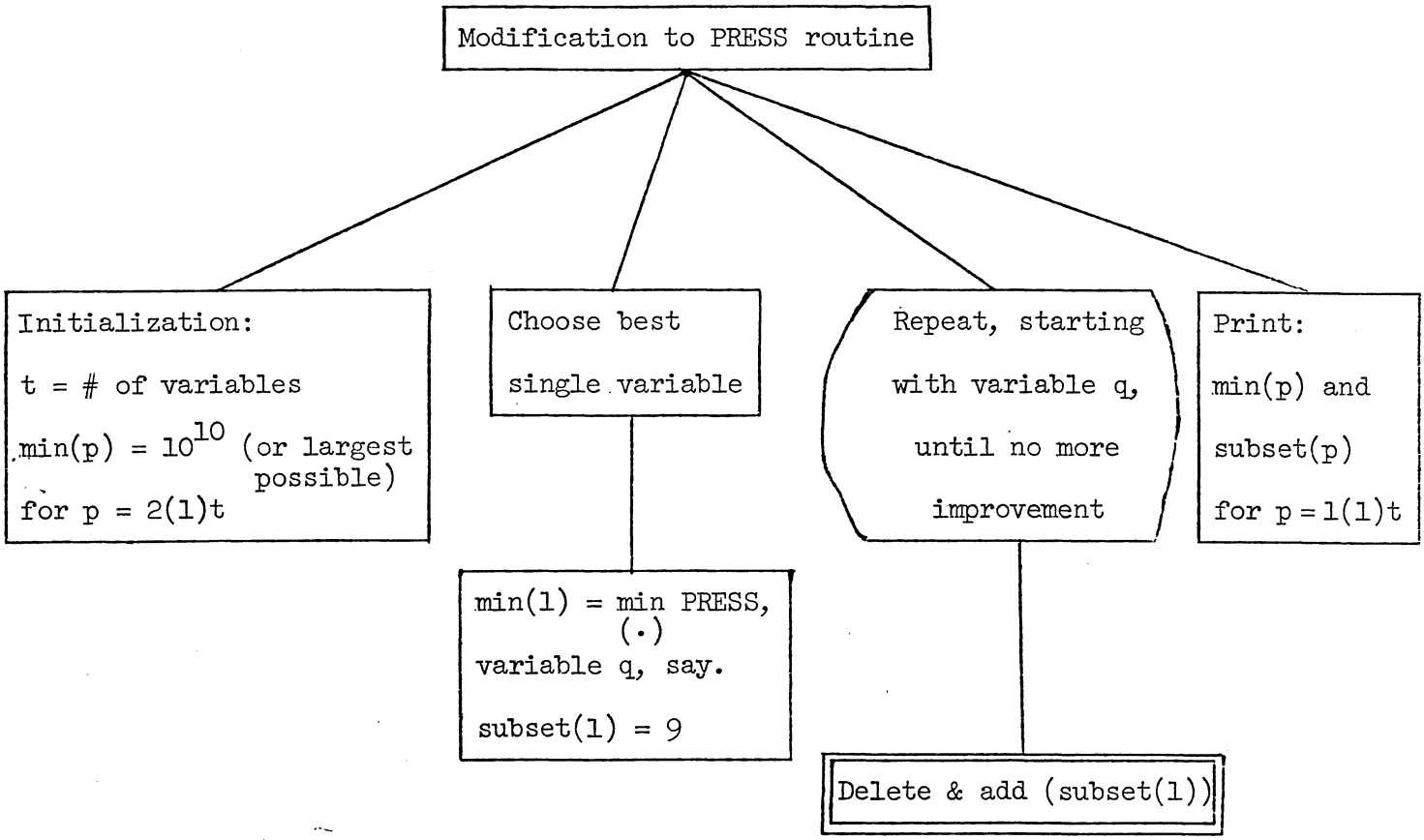
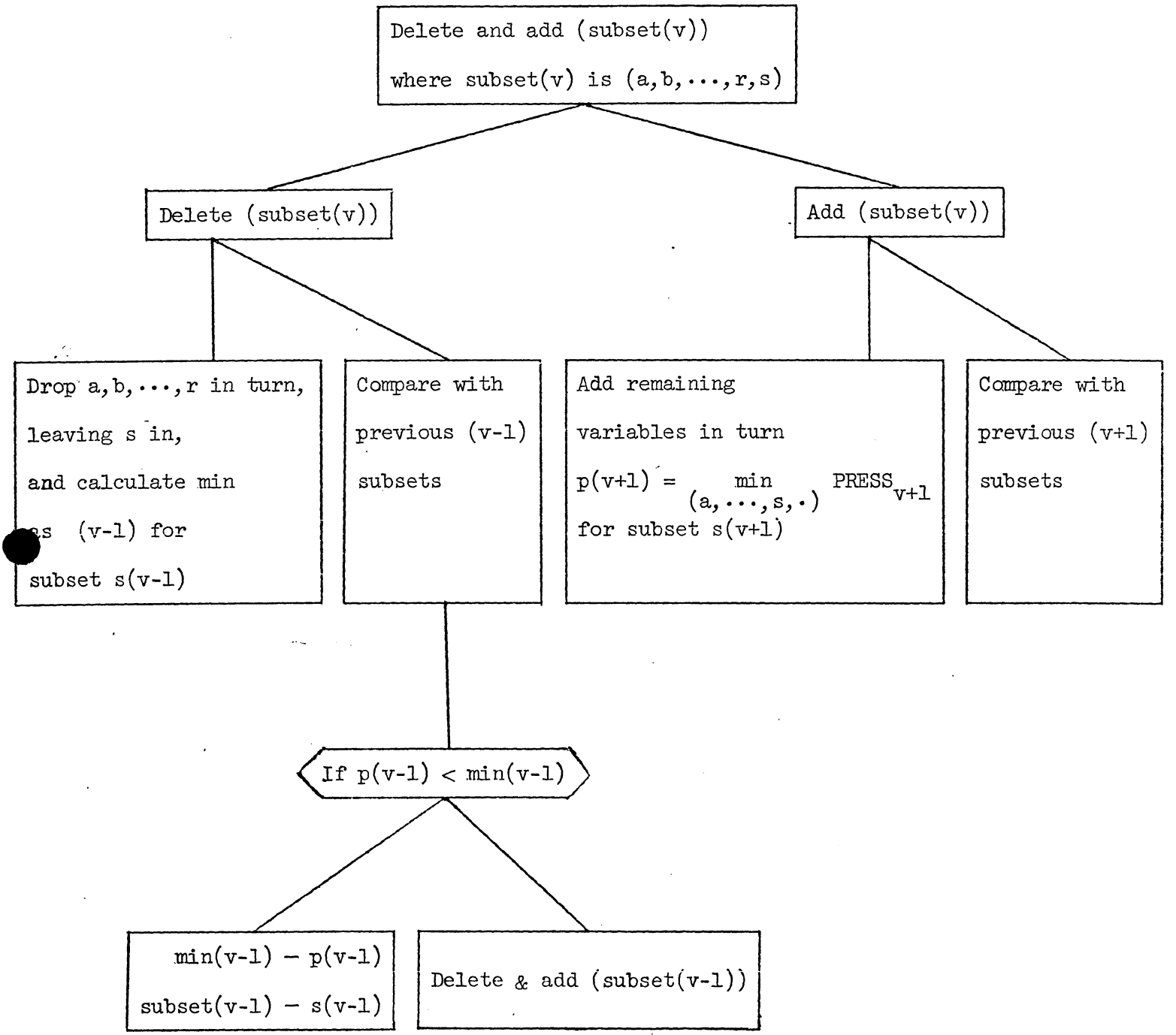


Figure 1



Note: 1) PRESS may be replaced by C_p or any other selection criteria.

2) Delete & add could be further modified to delete and add more than 1 variable at a time.

3) Printing press values at each stage could be included.

4) To prevent possible cycling could flag subsets already considered.