Variable Screening

MATH3805 Regression Analysis

Department of Mathematics

Hong Kong Baptist University

Fall 2021

Variable Screening

The purpose of variable screening methods is to determine quantitatively which candidate variables in a given list of available variables should be included as independent variables for explaining the variation in Y and which variables do not contribute to the variation in Y so that we do not include them in the model.

Obviously, we in fact have a dilemma.

- On one hand we would like to include as many of x_i as possible to make the model useful for prediction.
- On the other hand we should include as few of x_i as possible to make the model parsimonious.

This chapter is talking about variable selection, while the last chapter is talking about model building.

In real applications, we select variables first and then build a model based on the selected variables.

Thus, it is quite common that in the variable selection we consider only the first-order terms.

The first kind of variable selection methods we discussed is the **stepwise-type regression**, which includes (a) **forward selection**, (b) **backward elimination** and (c) **stepwise regression**.

(a) Forward selection:

Step 1. No independent variable in the model.

Step 2. Include in the model the independent variable (from the candidate variables that have not yet included in the model) that has the largest partial F-statistic (for testing whether the reduced model [the one without the new variable] is as good as the “complete” model [referring just to the one with the new variable, not really the true complete model including all candidate variables]), provided that its F -statistic is larger than F_{in} (F -to-enter). (This partial F-statistic will be just the square of the t-statistic for testing whether the parameter is zero. However, talking about F-statistic allows us to generalise this procedure to a procedure that could add more than one variable in this step, e.g. several dummy variables corresponding to one categorical independent variable.)

Step 3. Repeat Step 2 until the largest partial F-statistic among the remaining variables does not exceed F_{in} or when the last candidate variable is added.

Choosing the cutoff value F_{in} can be thought of as specifying a stopping criterion for this algorithm. Some computer programs allow the analyst to specify this number directly while some others (like SAS) require the choice of a type I error rate α_{in} to generate F_{in}.

That is to say, in step 2 we may specify a level-to-enter α_{in} such that the candidate variable with the smallest *p*-value is added to the model, provided that the *p*-value is less than α_{in}.

The variable having the largest partial *F*-statistic of course is the variable having the smallest *p*-value. However, these two stopping criteria (using a fixed F_{in} and using a fixed α_{in}) are not equivalent since the value of the degrees of freedom of the partial *F*-statistic depends on the number of parameters in the model, and hence is not a constant. That is, a fixed α_{in} corresponds to different F_{in} values in different steps.

Nowadays we will find that using a fixed α_{in} is a more natural stopping criterion.

However, in the old days when statisticians had been using tables and pocket calculators, calculating the *p*-values would be very laborious, and so using a fixed F_{in} allowed a more efficient algorithm.

It is anyway a screening procedure, involving many partial F-tests. Thus, a very precise significance level at each step of the selection process does not mean a well-controlled overall significance level for the resultant model. Thus, even nowadays we may use a fixed F_{in}.

(b) Backward elimination:

Step 1. Include all candidate variables in the model.

Step 2. Remove from the model the independent variable that has the smallest partial F-statistic, provided that its F-statistic is less than F_{out} (F-to-remove). [Alternatively but not equivalently, remove the independent variable that has the largest p-value, provided that its p-value is larger than α_{out}.]

Step 3. Repeat Step 2 until the smallest partial F among the existing independent variables in the model is not less than F_{out} [alternatively, the largest p-value is not larger than α_{out}] or when all independent variables are removed.

The backward elimination is particularly favoured by analysts who like to see the effect of including all the candidate variables, just so that nothing “obvious” will be missed. The disadvantage is that we have to fit models with many parameters in early steps, i.e. we have to pay a higher computational cost.

The two procedures above suggest a number of possible combinations. One of the most popular is the stepwise regression, a modification of forward selection in which at each step, after a variable entered, all independent variables now in the model are reassessed via their partial F-statistics. An independent variable added at an earlier step may now be redundant because of the relationships between it and other independent variables now in the model. If the smallest partial F-statistic among the existing independent variables is less that F_{out}, the corresponding independent variable is dropped from the model. [Alternatively, one may specify α_{in} and α_{out} instead of F_{in} and F_{out} and proceed in the same way.]

(c) Stepwise regression:

Step 1. No independent variable in the model.

Step 2. Include in the model the independent variable that has the largest partial F-statistic, provided that its F -statistic is larger than F_{in}. [Alternatively but not equivalently, include in the model the independent variable having the smallest p-value, provided that it is less than α_{in} when added to the model.] Fit the model with the new independent variable and remove from the model the independent variable that has the smallest partial F-statistic, provided that its F-statistic is less than F_{out}. [Alternatively but not equivalently, remove from the model the predictor having the largest p-value, provided that it is greater than α_{out}.]

Step 3. Repeat Step 2 until the largest partial F-statistic among the remaining variables does not exceed F_{in} [Alternatively but not equivalently, the smallest p-value among the remaining candidate predictors is not less than α_{in}.], or when all predictors are added and cannot be removed.

The stepwise regression procedure requires two cutoff values F_{in} and F_{out}.

Some analysts prefer to choose $F_{\text{in}} = F_{\text{out}}$, although it is not necessary.

If we choose $F_{\text{in}} > F_{\text{out}}$, we make it relatively more difficult to add an independent variable than to delete one.

We should never use $F_{\text{in}} < F_{\text{out}}$ (i.e. never use $\alpha_{\text{in}} > \alpha_{\text{out}}$) because a partial F-statistic lying between these two cutoff values will lead to cycling, where a variable is continually entered and removed.

It is popular to use $F_{\text{in}} = F_{\text{out}} = 4$, which corresponds roughly to (but not the same as) $\alpha_{\text{in}} = \alpha_{\text{out}} = 0.05$.

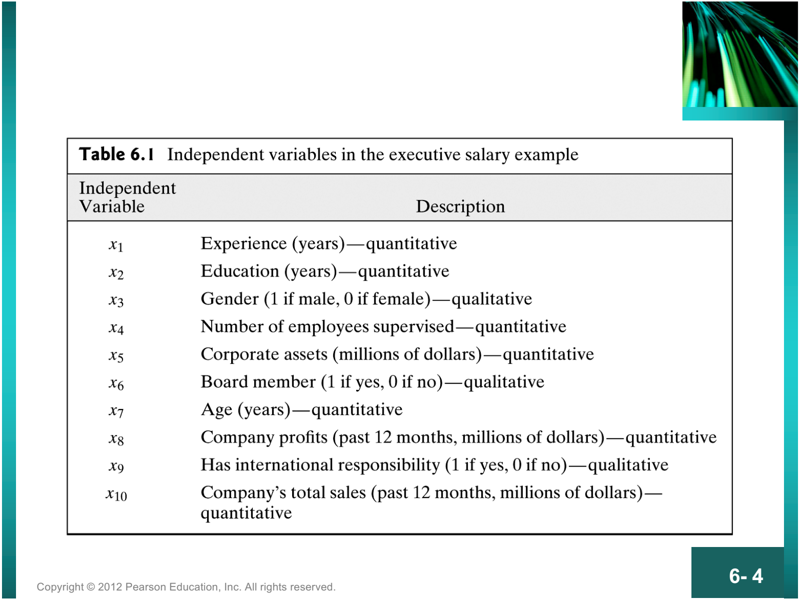
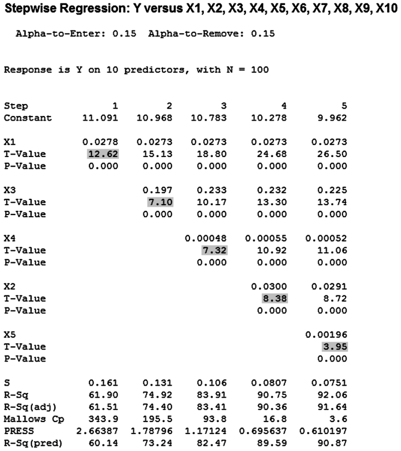


Table 6.1 Independent variables in the executive salary example

Independent Variable	Description
x_1	Experience (years)—quantitative
x_2	Education (years)—quantitative
x_3	Gender (1 if male, 0 if female)—qualitative
x_4	Number of employees supervised—quantitative
x_5	Corporate assets (millions of dollars)—quantitative
x_6	Board member (1 if yes, 0 if no)—qualitative
x_7	Age (years)—quantitative
x_8	Company profits (past 12 months, millions of dollars)—quantitative
x_9	Has international responsibility (1 if yes, 0 if no)—qualitative
x_{10}	Company's total sales (past 12 months, millions of dollars)—quantitative

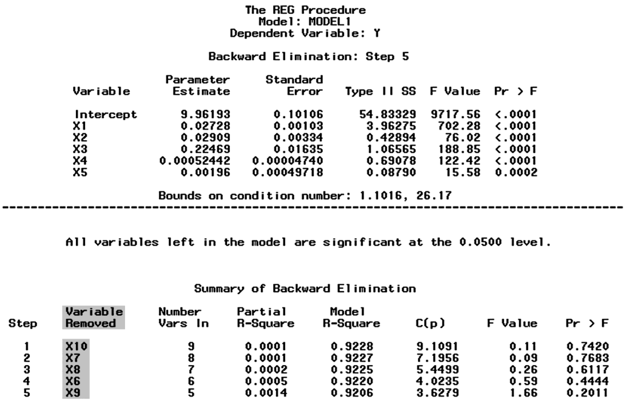
Figure 6.1 MINITAB stepwise regression results for executive salaries



Stepwise Regression: Y versus X1, X2, X3, X4, X5, X6, X7, X8, X9, X10

```
 Alpha-to-Enter: 0.15  Alpha-to-Remove: 0.15


Response is Y on 10 predictors, with N = 100


Step              1        2        3          4          5
Constant     11.091   10.968   10.783     10.278      9.962

X1           0.0278   0.0273   0.0273     0.0273     0.0273
T-Value       12.62    15.13    18.80      24.68      26.50
P-Value       0.000    0.000    0.000      0.000      0.000

X3                     0.197    0.233      0.232      0.225
T-Value                 7.10    10.17      13.30      13.74
P-Value                0.000    0.000      0.000      0.000

X4                            0.00048    0.00055    0.00052
T-Value                          7.32      10.92      11.06
P-Value                         0.000      0.000      0.000

X2                                        0.0300     0.0291
T-Value                                     8.38       8.72
P-Value                                    0.000      0.000

X5                                                  0.00196
T-Value                                                3.95
P-Value                                               0.000

S             0.161    0.131    0.106     0.0807     0.0751
R-Sq          61.90    74.92    83.91      90.75      92.06
R-Sq(adj)     61.51    74.40    83.41      90.36      91.64
Mallows Cp    343.9    195.5     93.8       16.8        3.6
PRESS       2.66387  1.78796  1.17124   0.695637   0.610197
R-Sq(pred)    60.14    73.24    82.47      89.59      90.87
```

Figure 6.2 SAS backward stepwise regression for executive salaries



```
                          The REG Procedure
                           Model: MODEL1
                       Dependent Variable: Y

                    Backward Elimination: Step 5

                 Parameter     Standard
   Variable       Estimate        Error    Type II SS   F Value   Pr > F

   Intercept       9.96193      0.10106      54.83329   9717.56   <.0001
   X1              0.02728      0.00103       3.96275    702.28   <.0001
   X2              0.02909      0.00334       0.42894     76.02   <.0001
   X3              0.22469      0.01635       1.06565    188.85   <.0001
   X4           0.00052442   0.00004740       0.69078    122.42   <.0001
   X5              0.00196   0.00049718       0.08790     15.58   0.0002

                 Bounds on condition number: 1.1016, 26.17
-----------------------------------------------------------------------------


          All variables left in the model are significant at the 0.0500 level.


                      Summary of Backward Elimination

         Variable     Number    Partial     Model
  Step   Removed      Vars In   R-Square   R-Square      C(p)    F Value   Pr > F

   1     X10             9       0.0001     0.9228     9.1091     0.11    0.7420
   2     X7              8       0.0001     0.9227     7.1956     0.09    0.7683
   3     X8              7       0.0002     0.9225     5.4499     0.26    0.6117
   4     X6              6       0.0005     0.9220     4.0235     0.59    0.4444
   5     X9              5       0.0014     0.9206     3.6279     1.66    0.2011
```

To conclude the stepwise-type procedures, we have to note the following comments:

1. It is NOT true that all important variables have been identified and unimportant variables removed.
2. A large number of single β-parameter t-test have been done, meaning that very probable that we included some unimportant variable (type I errors) and eliminate some important ones (type II errors).
3. It is likely not that there is one best subset model (a subset model means a model containing a subset of the candidate variables), but that there are several equally good ones.
4. The order in which the independent variables enter or leave the model does not necessarily imply an order of importance.
5. The three procedures do not necessarily lead to the same choice of final model.

6. The partial F value examined at each stage is the maximum of several correlated partial F-statistics (i.e. it is a result of multiple testing), thinking of p-value reported in the computer output as a level of significance or type I error rate of one single test for an individual parameter is misleading.
7. When we choose the variables to be included in the list of candidate variables, we may often omit higher order terms to keep the number of variables manageable. Thus, it is just a variable screening procedure and after we decide which independent variables have important main effect terms, we should then consider their second-order terms and other interactions, as we did in Chapter 5.
8. Stepwise-type regression should be used only when necessary, that is, when you want to determine which of a large number of potentially important independent variables should be used in the model-building process.

9. In the forward selection, the *MSE* (i.e. the sample variance $\hat{\sigma}^2$ of the residuals of the model) values will tend to be inflated during the initial steps, because important independent variables have been omitted. When they are omitted, the variation in the residuals comes not only from the random fluctuation caused by ε but also from the variation in the missing variables. (Variation in an independent variable is accompanied by variation in the response, even without the random error.) Thus, the sample variance of the residuals overestimates σ^2. This in turn leads to partial F-statistics that are too small (or p-values too large), making remaining candidate variables difficult to enter. In the backward elimination, the *MSE* values tend to be more nearly unbiased because important independent variables are retained at each step.

10. **NEVER** let a computer select independent variables mechanically. The computer does not know your research questions nor the literature upon which they rest. It cannot distinguish independent variables of direct substantive interest from those whose effects you want to control. Computer's job is to compute and we human beings are responsible for making decisions.

All-possible-regression selection procedure.
The procedure is to consider first the null model

$$Y = \beta_0 + \varepsilon,$$

then all one-variable models [there are in total $\binom{k}{1}$ such models]

$$Y = \beta_0 + \beta_i x_i + \varepsilon, \quad \text{for } i = 1, \ldots, k,$$

then all two-variable models [there are in total $\binom{k}{2}$ such models]

$$Y = \beta_0 + \beta_i x_i + \beta_j x_j + \varepsilon, \quad \text{for } i \neq j,$$

and so on, up to the full model

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon.$$

That is to say, we consider $\binom{k}{0} + \binom{k}{1} + \cdots + \binom{k}{k} = 2^k$ models. (Note that when e.g. $k = 10$, we have to consider 1024 models, and when $k = 13$, there are $2^k = 8182$ possible models.) Then choose one from all possible models, according to some criterion, as the “best” model.

Commonly used criteria.

1. R^2-criterion: find a subset model so that adding more variables to the model will yield only small increases in R^2. Unlike that in stepwise-type regression, the decision about when to stop adding variable is a subjective one. In general, the best models with p_1 parameters are not necessarily nested with the best models with p_2 parameters, where $p_2 > p_1$. (The phenomenon is not desirable and may cause interpretation problems.)

2. R_a^2-criterion/MSE-criterion: choose the model with the highest R_a^2 or choose the model by using the same graphical approach as the R^2 criterion except that we consider R_a^2 instead of R^2. Note that

$$R_a^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE_p}{SST_{\text{complete}}} = 1 - (n-1)\frac{MSE_p}{SST_{\text{complete}}},$$

where SSE_p is the error sum of squares of a p-parameter model and SST_{complete} is the total sum of squares of the complete model containing all candidate variables (as a matter of fact, $SST_{\text{complete}} = \sum(Y_i - \overline{Y})^2$ has nothing to do with the number of parameters in the model), and so the model of the highest R_a^2 is also the model of the lowest MSE_p. However, since $MSE_p = SSE_p/(n-p)$, having the lowest MSE_p among all models with different p does not necessarily imply the model also has the lowest SSE_p, unless we restrict to a fixed p.

3. Mallows' C_p-criterion: find the model such that (i) for prediction purpose, C_p close to or below the line $C_p = p$, or (ii) for parameter estimation purpose, C_p close to or below the line $C_p = 2p - k$, where k is the total number of candidate variables available. The former line $C_p = p$ was proposed by Mallows in 1973, while the latter line $C_p = 2p - k$ was suggested by Hocking in 1976. The idea of this C_p criterion is as follows. We define the total mean squared error (*TMSE*) by

$$TMSE := \sum \mathbb{E}\left\{[\hat{Y}_i - \mathbb{E}(Y_i)]^2\right\},$$

where $\hat{Y}_i$ is the fitted value of Y_i by the model under consideration, and we want to have a model whose *TMSE* is small. However, the value of *TMSE* also depends on the error variance σ^2. To standardise, we take the ratio $TMSE/\sigma^2$ but neither *TMSE* nor σ^2 can be calculated from the data. A good estimator of this ratio is Mallows' C_p, defined by

$$C_p := \frac{SSE_p}{MSE_{\text{complete}}} - (n - 2p) = (n - p)\frac{MSE_p}{MSE_{\text{complete}}} - (n - 2p).$$

(Note that in the textbook p denotes the number of independent variables and so the formula looks different from ours.) If MSE_P (the sample variance of the residuals $Y_i - \hat{Y}_i$, where $\hat{Y}_i$ comes from a model with p parameters) is equal to MSE_{complete}, then $C_p = p$, and typically (but empirically it is not always that) the variance of the residuals in a subset model is greater than that in the complete model, i.e. $MSE_p > MSE_{\text{complete}}$. Thus, it is desirable that the C_p value is close to or below the reference line $C_p = p$. The argument for the reference line $C_p = 2p - k$ is less obvious, and we do not go into details of this argument here and accept this as a rule of thumb.

4. *PRESS*-criterion: use the same graphical approach as the R^2 criterion, except that now we should look for small values for *PRESS*. However, SAS does not have this option available in the choices of selection method.

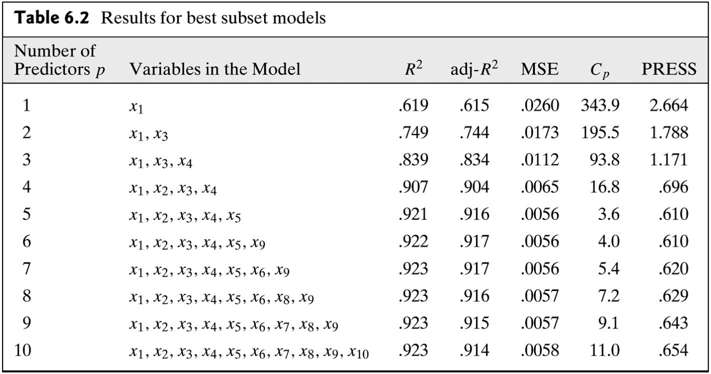
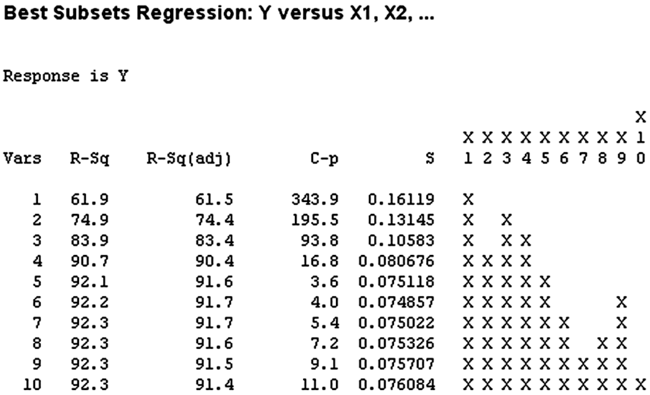


Table 6.2 Results for best subset models

Number of Predictors p	Variables in the Model	R^2	adj-R^2	MSE	C_p	PRESS
1	x_1	.619	.615	.0260	343.9	2.664
2	x_1, x_3	.749	.744	.0173	195.5	1.788
3	x_1, x_3, x_4	.839	.834	.0112	93.8	1.171
4	x_1, x_2, x_3, x_4	.907	.904	.0065	16.8	.696
5	x_1, x_2, x_3, x_4, x_5	.921	.916	.0056	3.6	.610
6	$x_1, x_2, x_3, x_4, x_5, x_9$	.922	.917	.0056	4.0	.610
7	$x_1, x_2, x_3, x_4, x_5, x_6, x_9$	.923	.917	.0056	5.4	.620
8	$x_1, x_2, x_3, x_4, x_5, x_6, x_8, x_9$	.923	.916	.0057	7.2	.629
9	$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9$	.923	.915	.0057	9.1	.643
10	$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$	.923	.914	.0058	11.0	.654

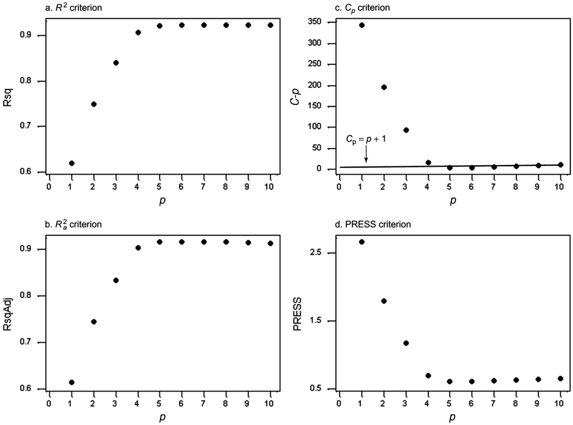
Figure 6.3 MINITAB all-possible-regressions selection results for executive salaries



Best Subsets Regression: Y versus X1, X2, ...

```
Response is Y

                                                                  X
                                                X X X X X X X X X 1
Vars   R-Sq    R-Sq(adj)        C-p         S   1 2 3 4 5 6 7 8 9 0

   1   61.9         61.5      343.9   0.16119   X
   2   74.9         74.4      195.5   0.13145   X   X
   3   83.9         83.4       93.8   0.10583   X   X X
   4   90.7         90.4       16.8  0.080676   X X X X
   5   92.1         91.6        3.6  0.075118   X X X X X
   6   92.2         91.7        4.0  0.074857   X X X X X       X
   7   92.3         91.7        5.4  0.075022   X X X X X X     X
   8   92.3         91.6        7.2  0.075326   X X X X X X   X X
   9   92.3         91.5        9.1  0.075707   X X X X X X X X X
  10   92.3         91.4       11.0  0.076084   X X X X X X X X X X
```

Figure 6.4 MINITAB plots of all-possible-regressions selection criteria for Example 6.2



a. R^2 criterion

b. R_a^2 criterion

c. C_p criterion

d. PRESS criterion

According to all four criteria, the variables $x_1,\ x_2,\ x_3,\ x_4$ and x_5 should be included in the group of the most important predictors.