Variable selection

The problem

- In many regression problems there will be:
  1. A fixed sample size to work with.
  2. A moderate to large number of potential predictor variables.
- Generally adding more variables to a regression model that already contains a small number of variables will improve predictive accuracy.

- Continuing to add variables (without adding more sample) will often lead to a deterioration in predictive accuracy (over-fitting).
- The goal is to find the *best* subset of variables: Bias/Variance trade-off.
- There are likely many subsets of variables that are likely to do well.
- Finding the best subset of variables is often referred as *variable selection*.

- For experiments variable selection is done before data collection as important variables are chosen for study based on theoretical aspects.
- For observational studies there is little control of how data is collected. In this case variable selection is done after date collection and in the data analysis domain.

Some effects of dropping variables

Assume that the correct model is given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in} + \varepsilon_i$$

and consider the sub-model which includes the first $p-1 < n$ independent variables:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{(p-1)} x_{i(p-1)} + \varepsilon_i.$$

- Deleting independent variable usually biases the estimates of the parameters left in the model;
- Deleting independent variable usually increases the values of the expectation of s^2 and decreases the covariance matrix of the estimates $\beta_0, \ldots, \beta_{(p-1)}$. Note that we are referring to the covariance matrix defined in terms of σ^2 and not its estimate s^2.

- A measure of the bias in the predicted values is called Mallow's C_p, where

$$C_p = \frac{\text{RSS}_p}{s^2} - (m - 2p).$$

and m is the sample size (number of observations), RSS_p are the residual sum of squares of the p-variable model and the estimator of σ^2 is given by:

$$s^2 = \frac{\text{RSS}_{n+1}}{(m - n - 1)}.$$

The key property for applications is that if the submodel does not lead to much bias in the predicted, then

$$C_p \approx p.$$

Notice that,

$$\begin{aligned} C_{n+1} &= \frac{\mathrm{RSS}_{n+1}}{s^2} - (m - 2(n+1)) \\ &= \frac{\mathrm{RSS}_{n+1}}{\mathrm{RSS}_{n+1}/(m-n-1)} - (m - 2(n+1)) \\ &= (m-n-1) - (m - 2(n+1)) \\ &= n+1. \end{aligned}$$

Thus, C_p assumes that the complete model has been carefully chosen so as to give reasonable assurance of negligible bias.

Effects on estimates of β_j

Assume that the correct model is given by

$$y = X\beta + \varepsilon, \quad \text{or} \quad y = (X_1 \quad X_2)\begin{pmatrix}\beta_1 \\ \beta_2\end{pmatrix} + \varepsilon,$$

where $\beta = (\beta_1 \quad \beta_2)^T$, $X = (X_1 \quad X_2)$, β_1 is a p-element vector and the other dimensions are chosen appropriately.

Thus, the correct model can be written as:

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon,$$

where $\mathrm{E}(\varepsilon) = 0$, $\mathrm{E}(y) = X_1\beta_1 + X_2\beta_2$ and $\mathrm{Var}(\varepsilon) = \sigma^2 I$.

Assume that we leave out $X_2\beta_2$ and obtain the sub-model:

$$y = X_1\beta_1 + \varepsilon.$$

The estimator of β_1 in the sub-model is given by:

$$\widehat{\beta}_1 = (X_1^T X_1)^{-1} X_1 y$$

and

$$\begin{aligned}\mathrm{E}(\widehat{\beta}_1) &= (X_1^T X_1)^{-1} X_1^T \mathrm{E}(y) \\ &= (X_1^T X_1)^{-1} X_1^T (X_1\beta_1 + X_2\beta_2) \\ &= \beta_1 + (X_1^T X_1)^{-1} X_1^T X_2\beta_2.\end{aligned}$$

Thus, the estimate of β_1 after deleting $X_2\beta_2$ is biased by

$$\boxed{\mathrm{E}(\widehat{\beta}_1) - \beta_1 = (X_1^T X_1)^{-1} X_1^T X_2\beta_2.}$$

The variance of β_1 in the sub-motel is given by:

$$\mathrm{Var}(\beta_1) = \sigma^2 (X_1^T X_1)^{-1}.$$

However, based on the full (correct) model:

$$\text{Var}(\beta) = \sigma^2 (X^T X)^{-1}$$

$$= \sigma^2 \begin{pmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{pmatrix}^{-1} \equiv \sigma^2 \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}$$

If $\beta^{(1)}$ corresponds to the first $(p-1)$ independent variables of the full model, then
$\text{Var}(\beta^{(1)}) = V_{11} \geq (X_1^T X_1)^{-1}$ and thus,

$$\boxed{\text{Var}(\beta_1) \leq \text{Var}(\beta^{(1)}).}$$

Variable selection procedures

When presented with a data set containing a large number of potential regressors, it can be a formidable task to identify a useful, well fitting and parsimonious regression model.

In the 1950s and 1960s statisticians spent a lot of time inventing ingenious ways of automating variable selection in multiple linear regression.

The four most popular methods are *forward selection*, *backward selection*, *stepwise regression* and *best subset regression*.

The idea of the first three is to reduce the number of possible models that need to be fitted.

Often these methods do not provide the same models which in most cases are not the optimal ones.

Here Optimal in the sense that there are better models to be selected using the same selection criteria.

The fourth procedure generates the best one-regressor model, the best two-regressor model, the best three-regressor model etc.

The best of these generated model is selected based on some criteria (R^2, adjusted R^2, C_p, etc).

Often an exhaustive search (considers all possible submodels) is used for generating all subset models.

Forward Selection

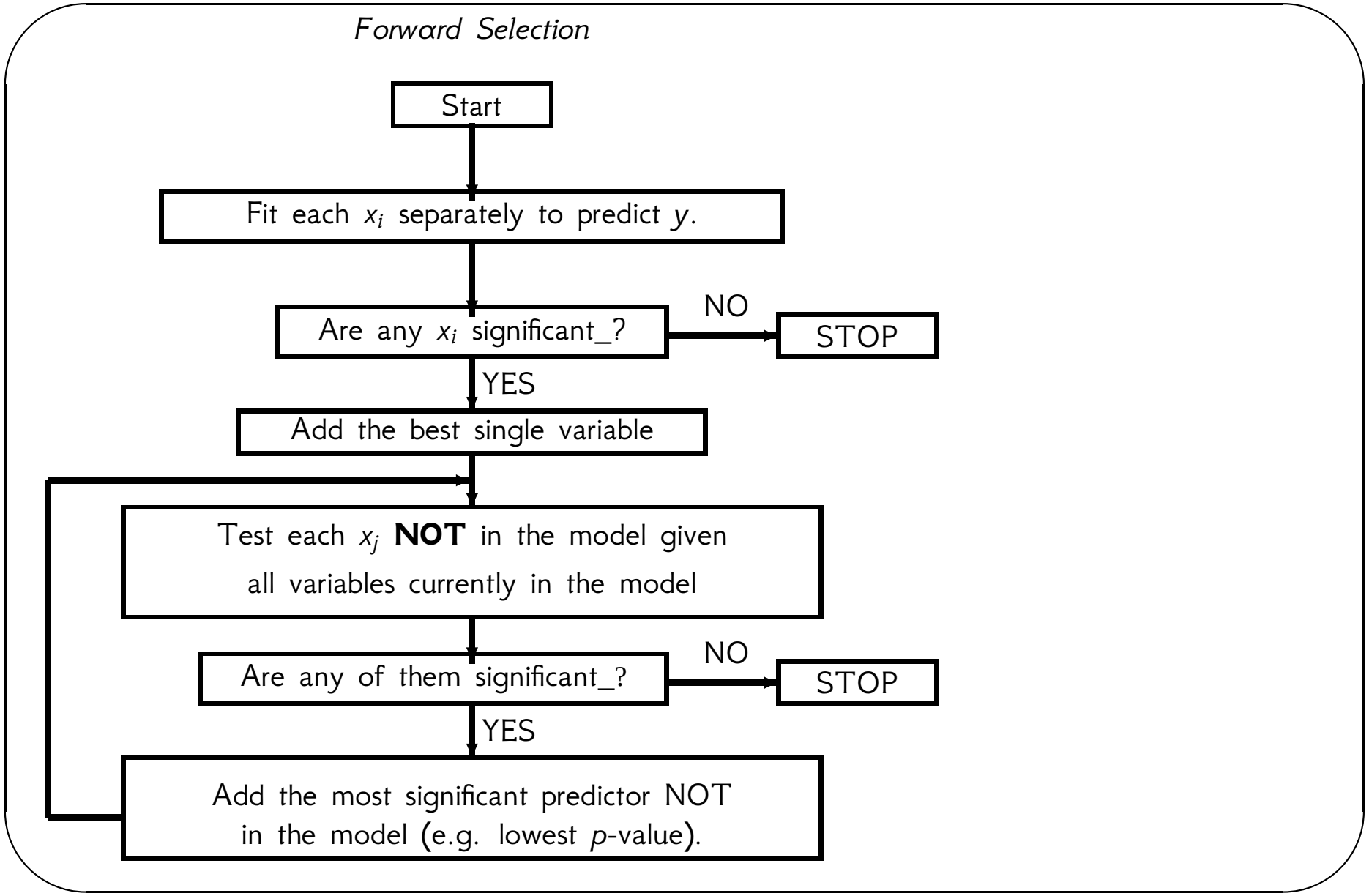
- Unlike exhaustive search, forward selection is always computationally tractable. Even in the worst case, it checks a much smaller number of subsets before finishing.
- This technique adds predictor variables and never deletes them.
- The starting subset in forward selection is the empty set.

- For a regression model with n possible predictor variables, the first step involves evaluating n predictor variable subsets, each consisting of a single predictor variable, and selecting the one with the highest evaluation criterion.

- The next step selects from among $(n-1)$ subsets, the next step from $(n-2)$ subsets, and so on.

- Even if all predictor variables are selected, at most $n(n+1)/2$ subsets are evaluated before the search ends.

1. Start with no variable.
2. For each variable NOT in the model, check the p-value if there are added in the model.
3. Choose the one with lowest p-value less than α.
4. Continue until no new variable can be added.


Forward Selection
Start
Fit each x_i separately to predict y.
Are any x_i significant_?
NO
STOP
YES
Add the best single variable
Test each x_j **NOT** in the model given all variables currently in the model
Are any of them significant_?
NO
STOP
YES
Add the most significant predictor NOT in the model (e.g. lowest p-value).

Example: Forward selection using the House data

Fit each variable separately	P-Value
Number of bedrooms	0.39
Floor space in square feet	0.000
Number of fireplaces	0.0475
Number of rooms	0.002
Storm windows (1 if present, 0 if absent)	0.019
Front footage of lot in feet	0.031
Annual Taxes	0.010
Number of bathrooms	0.003
Construction (0 if frame, 1 if brick)	0.504
Garaze Size (0 no car, 1 one car)	0.004
Condition (1 need work, 0 otherwise)	0.799
Location 1 (if property is in zone A, 0 otherwise)	0.010
Location 2 (if property is in zone B, 0 otherwise)	0.994

At 5% significance level the most significant variable is Floor space in square feet. This variable (FLR) is introduced to the model (and will never be deleted).

We try to find the second explanatory variable by examining their significance in the model which consists of FLR.

1)	P-Value	R2	Ad. R2
Floor space in square feet	0.000	0.557	0.518
Number of Bedrooms	0.388		

2)	P-Value	R2	Ad. R2
Floor space in square feet	0.000	0.543	0.503
Number of fireplaces	0.0427		

3)	P-Value	R2	Ad. R2
Floor space in square feet	0.000	0.544	0.505
Number of rooms	0.707		

4)	P-Value	R2	Ad. R2
Floor space in square feet	0.000	0.675	0.646
Storm windows	0.005		

5)	P-Value	R2	Ad. R2
Floor space in square feet	0.000	0.575	0.538
Front footage of lot in feet	0.194		

6)	P-Value	R2	Ad. R2
Floor space in square feet	0.000	0.544	0.504
Annual Taxes	0.726		

7)	P-Value	R2	Ad. R2
Floor space in square feet	0.000	0.580	0.543
Number of bathrooms	0.162		

```
8)                                P-Value    R2    Ad. R2
Floor space in square feet         0.000    0.609   0.575
Construction                       0.059

9)                                P-Value    R2    Ad. R2
Floor space in square feet         0.000   0.614   0.580
Garaze size                        0.049

10)                               P-Value    R2    Ad. R2
Floor space in square feet         0.000   0.542   0.502
Condition                          0.990

11)                               P-Value    R2    Ad. R2
Floor space in square feet         0.000   0.567   0.530
Location (zone A)                  0.254

12)                               P-Value    R2    Ad. R2
Floor space in square feet         0.000   0.556   0.517
Location (zone B)                  0.403
```

At 5% significance level the second most significant explanatory variable in the model (conditional to the FLR variable) is Storm windows (ST).

This variable enters the model which now consists of the constant, FLR and ST.

The remaining variables are fitted to this model one at a time. The most significant (if any) will enter the model.

```
        p-value      p-value          p-value
1) FLR  0.000    2) FLR 0.000     3) FLR 0.001
   ST   0.002       ST  0.001        ST  0.007
   BDR  0.099       FP  0.011        RMS 0.854
Adj R-squ: 0.67  Adj R-squ: 0.73  Adj R-squ: 0.63
```

```
4) FLR  0.000     5) FLR  0.000     6) FLR  0.000
   ST   0.003        ST   0.015        ST   0.006
   LOT  0.069        BTH  0.526        CON  0.062
Adj R-squ: 0.68   Adj R-squ: 0.64   Adj R-squ: 0.69

7) FLR 0.000   8) FLR 0.000  9) FLR  0.000  10) FLR 0.000
   ST  0.007      ST  0.006     ST   0.012      ST  0.006
   GAR 0.058      GDN 0.671     L1   0.625      L2  0.318
Adj  R2: 0.64   Adj R2: 0.63  Adj R2: 0.63  Adj R2: 0.65
```

The variable FP enters the model.

The remaining variables are fitted one at a time to the model comprising (FLR, ST, FP).

Their p-values are:

BDR (0.027), RMS (0.593), LOT (0.272), TAX (0.823), BTH (0.565), CON (0.149), GAR (0.265), GDN (0.634), L1 (0.979), L2 (0.222).

The variable BDR enters the model which now comprises (FLR, ST, FP, BDR).

Against this model the remaining variables have p-values:

RMS (0.009), LOT (0.053), TAX (0.157), BTH (0.075), CON (0.513), GAR (0.351), GDN (0.906), L1 (0.798), L2(0.187).

The variable RMS enters the model which now comprises (FLR, ST, FP, BDR, RMS).

Against this model the remaining variables have p-values:

LOT (0.062), TAX (0.697), BTH (0.218), CON (0.735), GAR (0.338), GDN (0.388), L1 (0.951), L2 (0.362)

None of this variables are significant at 5% and thus, the previous model is kept. That is,

```
Coefficients:
            Estimate    SE  t value   Pr(>|t|)
(Intercept) 24.172   4.901   4.93   8.1e-05
FLR          0.019   0.003   5.72   1.3e-05
ST          11.253   2.346   4.80   0.00011
FP          10.295   2.850   3.61   0.00174
BDR         -7.827   1.978  -3.96   0.00078
RMS          4.864   1.672   2.91   0.00868
---
Residual standard error: 5.132 on 20 DF
Multiple R-Squared: 0.87, Adjusted R-squared: 0.83
F-statistic: 26.05 on 5 and 20 DF, p-value: 4.1e-08
```

Backward Selection

- Backward selection has computational properties that are similar to forward selection. The starting subset in backward selection includes all possible predictor variables.
- Predictor variables are deleted one at a time as long as this results in a subset with a higher evaluation criterion.
- Again, in the worst case, at most $n(n+1)/2$ subsets must be evaluated before the search ends. Like forward selection, backward selection is not guaranteed to find the subset with the highest evaluation criterion.

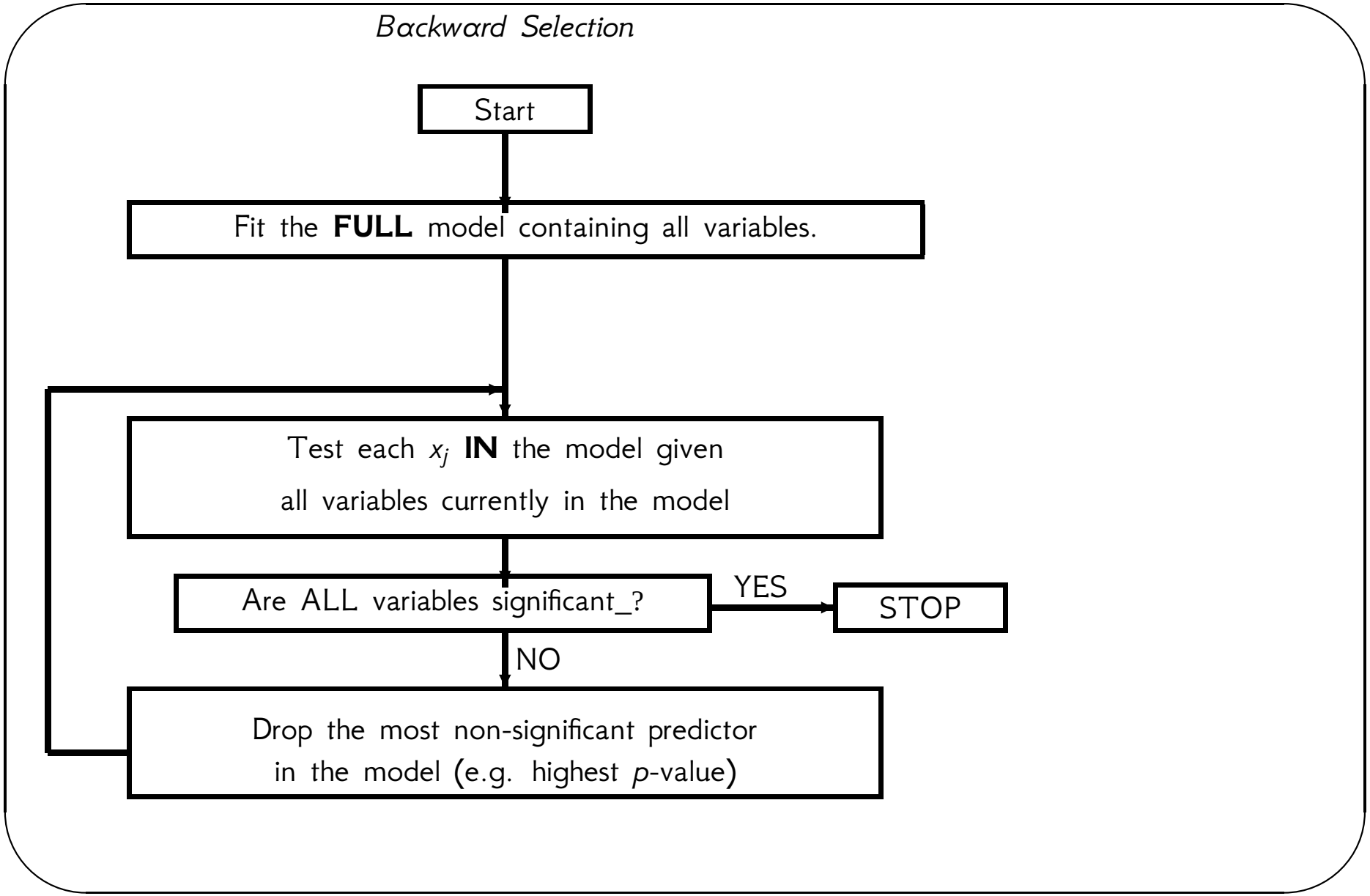
- The disadvantage of backward selection is that one's confidence in subset evaluation criterion values tends to be lower than with forward selection.

  This is especially true when the number of rows in the predictor matrix is close to the number of possible predictor variables.

  In such a case, there are very few points that the regression model can use in order to determine its parameter values, and the function evaluation criterion will be sensitive to small changes to the predictor matrix data.

- When the ratio of predictor matrix rows to predictor variables is small, it is usually a better idea to use forward selection than backward selection.

1. Start with all variables in the model.
2. Remove the variable with the highest p-value greater than α.
3. Refit the model and go to step 2.
4. Stop when ALL p-values are less than α.


Backward Selection
Start
Fit the **FULL** model containing all variables.
Test each x_j **IN** the model given
all variables currently in the model
Are ALL variables significant_?
YES
STOP
NO
Drop the most non-significant predictor
in the model (e.g. highest p-value)

Example: Backward selection using the House data

Fit the model comprising all variables

Variable	Coefficients	SE	t	Sig.
(Constant)	11.608	6.469	1.794	0.098
num of bedrooms	-4.502	2.243	-2.007	0.068
floor space	1.583E-02	0.006	2.641	0.022
num of fireplaces	-2.139	2.930	-0.730	0.479
number of rooms	2.380	1.911	1.246	0.237
storm windows	8.931	2.430	3.675	0.003
front footage	0.385	0.130	2.953	0.012
annual taxes	2.344E-04	0.005	0.051	0.960
num of bathrooms	1.187	2.849	0.417	0.684
construction	4.625	2.387	1.938	0.077
garaze size	4.985	1.514	3.292	0.006
condition	-0.212	2.653	-0.080	0.938
location(zone A)	1.565	3.132	0.500	0.626
location(zone B)	7.383	2.953	2.500	0.028

R2=0.936 Adj R2=0.867 s=

The less significant variable is the *Annual taxes*. This variable is deleted from the model (and never re-considered) and the model is fitted again.

Fit the model with all variables excluding Anual Taxes

Variable	Coefficients	SE	t	Sig.
(Constant)	11.802	5.009	2.356	0.035
num of bedrooms	-4.524	2.115	-2.139	0.052
floor space	1.608E-02	0.003	4.984	0.000
num of fireplaces	-2.120	2.793	-0.759	0.461
number of rooms	2.353	1.760	1.337	0.204
storm windows	8.946	2.318	3.860	0.002
front footage	0.386	0.122	3.170	0.007
num of bathrooms	1.132	2.526	0.448	0.662
construction	4.596	2.226	2.065	0.059
garaze size	5.000	1.429	3.498	0.004
condition	-0.254	2.419	-0.105	0.918
location(zone A)	1.599	2.937	0.545	0.595
location(zone B)	7.419	2.757	2.691	0.019

R2=0.936 Adj R2=0.877 s=

The less significant variable is the *Condition*. This variable is deleted from the model (and never reconsidered) and the model is fitted again. This procedure is repeated until all the variables in the model are significant.

The model comprising only significant variables

Variable	Coefficients	SE	t	Sig.
(Constant)	16.268	4.019	4.048	0.001
num of bedrooms	-2.051	1.034	-1.983	0.063
floor space	1.865E-02	0.003	6.896	0.000
storm windows	9.910	2.024	4.897	0.000
front footage	0.406	0.111	3.648	0.002
construction	5.622	1.927	2.918	0.009
garaze size	5.503	1.242	4.432	0.000
location(zone A)	8.087	2.026	3.992	0.001

R2=0.919 Adj R2=0.888 s=

In this case the model is given by:

$$\textbf{Price} = 16.268 - 2.951\,\textbf{BDR} + 0.0187\,\textbf{FLR} + 9.910\,\textbf{ST} + 0.406\,\textbf{LOT} + 5.622\,\textbf{CON} + 5.503\,\textbf{GAR} + 8.087\,\textbf{L2}.$$

(*In the Backward method for illustration a 90% level has been used. If 95% level was used, then the number of bedrooms is insignificant in the last regression fit and thus, it should be deleted and continue with the process until all variables are found to be significant.*)

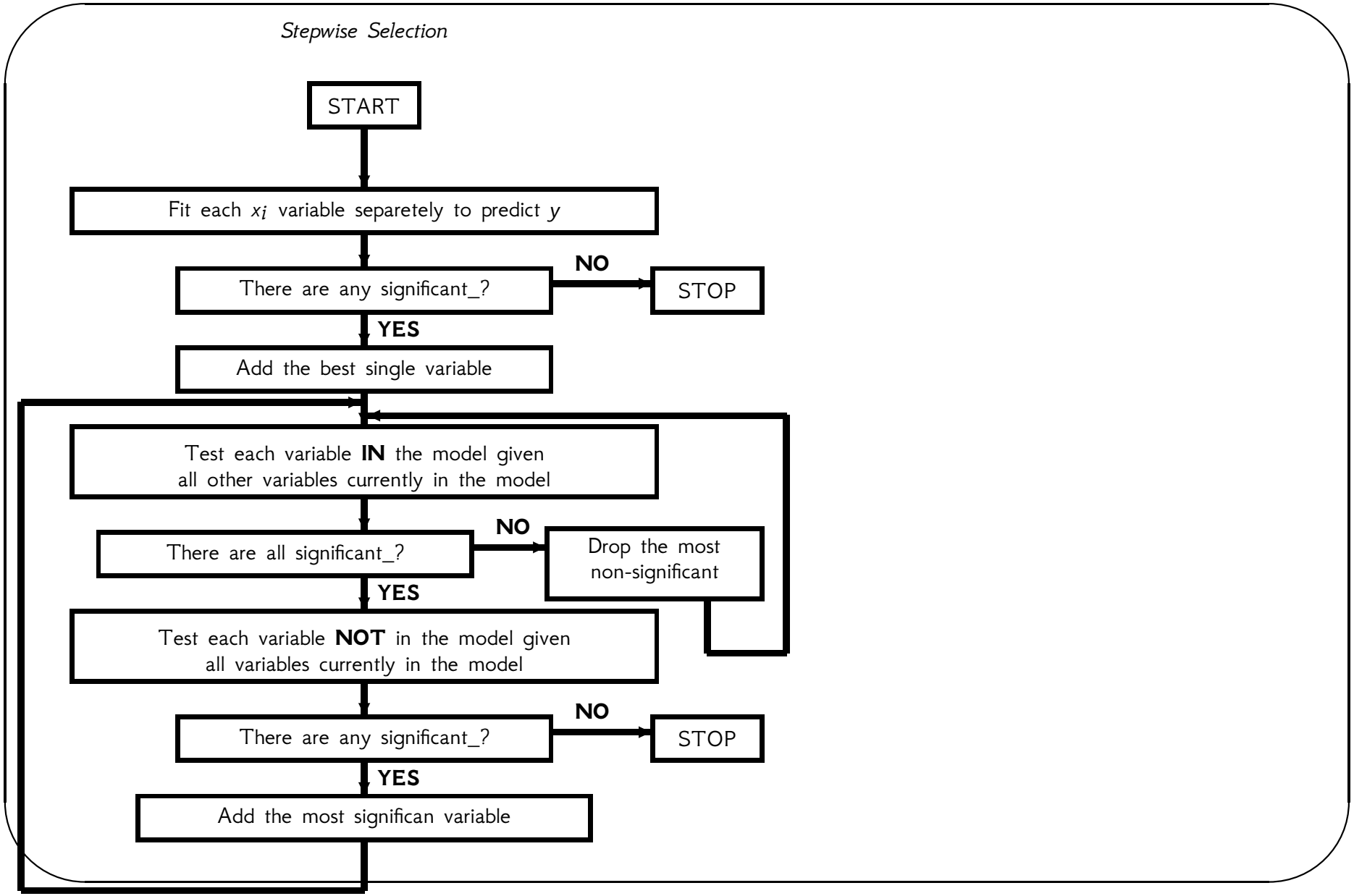
Stepwise Selection

- Stepwise selection has been proposed as a technique that combines advantages of forward and backward selection.
- At any point in the search, a single predictor variable may be added or deleted.
- Commonly, the starting subset is the empty set.
- At most n^2 subsets are evaluated before stepwise selection ends.

- There are, however, no guarantees that each predictor will be chosen at most one time.

- No strong theoretical results exist for comparing the effectiveness of stepwise selection against forward or backward selection.

- Stepwise selection evaluates more subsets than the other two techniques, so in practice it tends to produce better subsets Of course, the price that stepwise selection pays for finding better subsets is reduced computational speed: usually more subsets must be evaluated before the search ends.


Stepwise Selection
START
Fit each x_i variable separetely to predict y
There are any significant_?
NO
STOP
YES
Add the best single variable
Test each variable IN the model given all other variables currently in the model
There are all significant_?
NO
Drop the most non-significant
YES
Test each variable NOT in the model given all variables currently in the model
There are any significant_?
NO
STOP
YES
Add the most significan variable

For the *house data* it has been already observed that the most significant variable is the floor (FLR). This variable enters the model. The remaining variables are fitted one at time to the model comprising the constant term and the FLR. Again from the earlier results of the forward selection it is know that the strongest variable is the ST (Storm windows). Thus, this variable enters the model:

```
4)                                P-Value    R2     Ad. R2
Floor space in square feet         0.000   0.675   0.646
Storm windows                      0.005
```

Both variables (FLR and ST) are highly significant.

Thus, none of them is deleted. The remaining variables are now fitted one at a time to the model:

$$\textbf{Price} = 32.594 + 0.0189\,\textbf{FLR} + 10.226\,\textbf{ST}.$$

This gives:

```
1)                      P-Value
Floor in square feet    0.000
Storm windows           0.002
Number of bedrooms      0.099

2)                      P-Value
Floor in square feet    0.000
Storm windows           0.007
```

```
Number of fireplaces    0.903

3)                      P-Value
Floor in square feet    0.000
Storm windows           0.007
Number of rooms         0.854

4)                      P-Value
Floor in square feet    0.000
Storm windows           0.003
Front footage           0.069

5)                      P-Value
Floor in square feet    0.000
Storm windows           0.005
Annual Taxes            0.493
```

6)	P-Value
Floor in square feet	0.000
Storm windows	0.015
Number of bathrooms	0.526

7)	P-Value
Floor in square feet	0.000
Storm windows	0.006
Construction	0.062

8)	P-Value
Floor in square feet	0.000
Storm windows	0.007
Garage size	0.05

```
9)                      P-Value
Floor in square feet    0.000
Storm windows           0.006
Condition               0.671

10)                     P-Value
Floor in square feet    0.000
Storm windows           0.012
Location 1              0.625

11)                     P-Value
Floor in square feet    0.000
Storm windows           0.006
Location 2              0.318
```

At 5% significance none of the remaining variables are found to be significant. Thus, the procedure terminates and the previous selected model which comprises the variables FLR and ST is chosen. This is the same model derived using the forward selection method:

Variable	Coefficient	SE	t	Sig.
(Constant)	32.594	3.907	8.343	0.000
floor space	1.891E-02	0.003	5.740	0.000
storm windows	10.226	3.337	0.368	0.005

R2=.675 Adj. R2=.646 s=7.48354

$$\textbf{Price} = 32.594 + 0.0189\,\textbf{FLR} + 10.226\,\textbf{ST}.$$

Subset selection

- The search over all possible subsets of independent variables allow us to examine all regression equations constructed out of a given list of variables.
- A measure of fit for each regression is also generated. Such measure is the coefficient of determination R^2, adjusted R^2, C_p, etc.
- These measures are used in order to determine the ***best*** regression equation.
- *The procedure is the most useful if the number of*

variables is not too large.

- If there are n independent variables, then there are $2^n - 1$ possible combinations of subset variables. E.g. for the 3 variables x_1, x_2 and x_3 there are $2^3 - 1 = 7$ combinations:

$$(x_1), (x_2), (x_3), (x_1, x_2), (x_1, x_3), (x_2, x_3), (x_1, x_2, x_3).$$

- Consider the n-variable regression model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{2n} + \varepsilon_i.$$

There 2^n subset regression models that include the constant β_0. This includes the *null* regression $\mathrm{E}(y) = \beta_0$.

Generating all sub-models: infeasible

# Variables	Time	Sub-models
20	1 second	1 Million
26	1 minute	60 Million
32	1 hour	3.6 Billion
37	1+ day	86.5 Billion
41-42	1+ Month	2.6 Trillion
45	1 year	31 Trillion
52	1+ century	10^{15}
55	1 millennium	10^{16}
60	30+ millenniums	10^{18}

For not very small number of variables:

- It takes enormous time to generate all subset models.

  Thus, it is computationally infeasible to generate all submodels from 40+ variables.

- The number of models generated is extremely big.

  This makes it difficult to process (space requirement and analysis).

Regression Trees

- For the derivation of the submodels a ***Dropping Variable Method*** is employed. It is based on a ***regression tree*** and a ***dot*** ($\bullet$) notation.

- The variables are represented by their index. That is, $\mathbf{i}$ denotes the variable $\mathbf{x_i}$.

- A ***node*** comprises a sequence of indexes (variables) and a ***dot***.

- The submodels are given by the the variables (indexes) in the nodes without the *dot*.

  Thus, $(1,2,\bullet 5,6)$ denotes the submodel with variables x_1, x_2, x_5 and x_6, or equivalently the regression

$$\mathrm{E}(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_5 x_5 + \beta_6 x_6.$$

- The empty node $(\bullet)$ denotes the regression $\mathrm{E}(y) = \beta_0$.

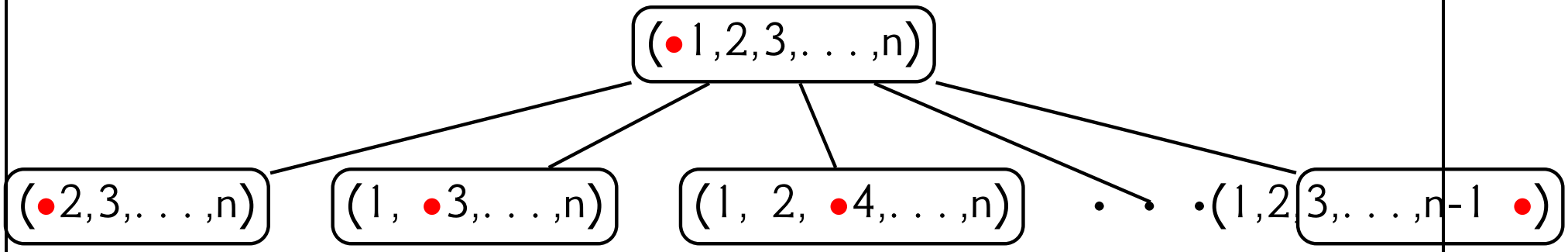
- The root node of the regression tree is given by:

$$(\bullet 1, 2, 3, \ldots, n).$$

- The root node has n children-node which are defined by replacing the **dot** with a variable (index) that appear in its right. That is, the root has the children-nodes:

$$(\bullet 2,3,\ldots,n),\ (1,\bullet 3,\ldots,n),$$
$$(1,2,\bullet 4,\ldots,n),\ \ldots,\ (1,2,3,\ldots,n-1\bullet).$$

This is denoted by the tree:

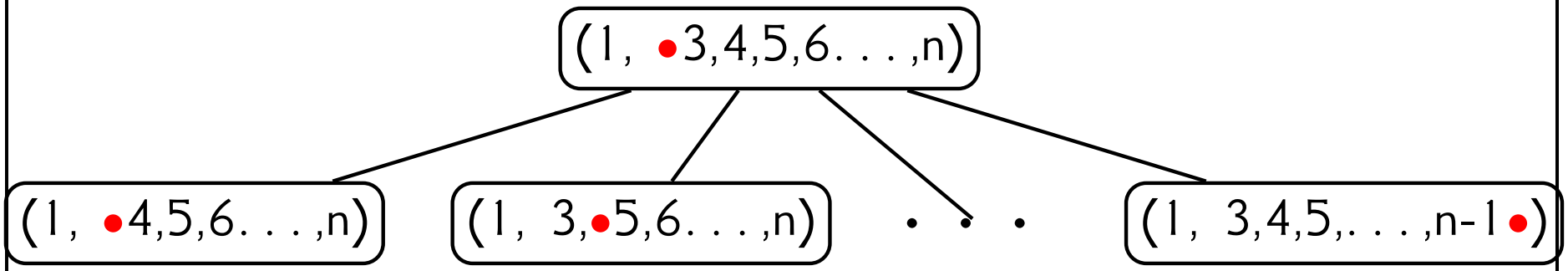


- Each node can generate further children-nodes by employing the same method (replacing the *dot* with the variable in its right), unless the *dot* is at the last position (since there are no variables in the right of the *dot*).

- For example, the node $(1, 2, 3, \ldots, n-1\bullet)$ has no children, while the node $(1, \bullet 3, 4, 5, 6 \ldots, n)$ has the children nodes:

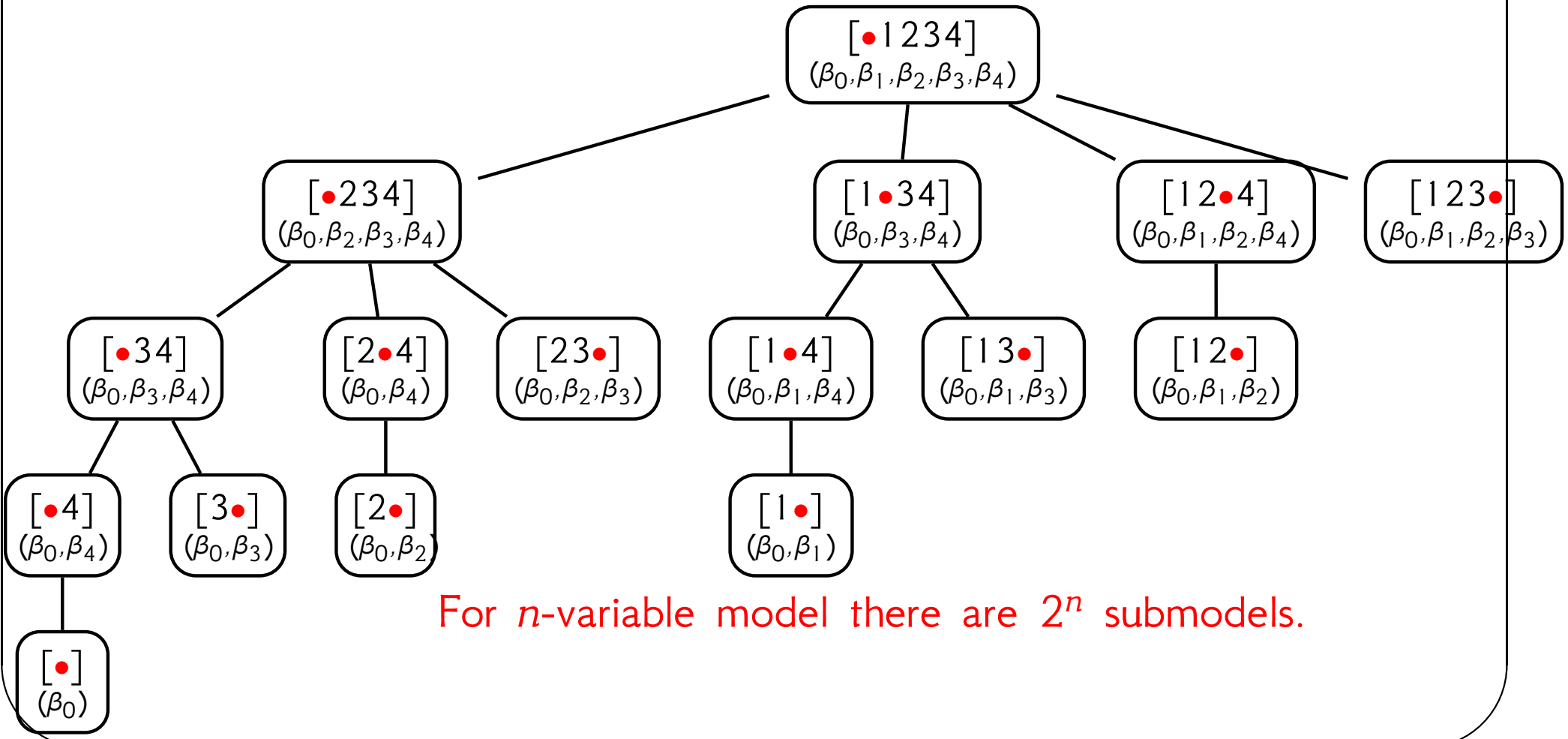
$$(1, \bullet 4, 5, 6 \ldots, n), \quad (1, 3, \bullet 5, 6 \ldots, n),$$
$$(1, 3, 4, \bullet 6, \ldots, n), \quad \ldots, \quad (1, 3, 4, 5, \ldots, n-1\bullet).$$

This is represented as a the tree:



- Note that the variables before the *dot* are included in children-nodes, i.e. in the subsequent generated models. In the example above the variable 1, i.e. x_1, is included in all generate children nodes.

The regression tree



For n-variable model there are 2^n submodels.

Generating the best subset models

- Regressions with more than 40 variables are not uncommon. Therefore a method that restricts the search to the more useful regressions is commonly considered.

- The search should at least find for each value $\nu = 1, 2, \ldots, n$ the best ν-variable regressor subset (including the constant β_0) which has the minimum $\text{RSS}_{\nu+1}$, or equivalently best $C_{\nu+1}$.

- It is often desirable to obtain the best few models (not only the best) for each variable.

- There are $\binom{n}{\nu} = n!/(\nu!(n-\nu)!)$ possible ν-variable models.
- The search of the *best* submodel should also generate the competitive submodels closed to the best one. In this way the decision of the most appropriate model could be taken (*Appropriate model implies a model with a C_p close to the best one and for which there is a theoretical or otherwise explanation*).

- An extension of the subset search method is to consider models comprising variables within a predefined range.

  That is, the search should find the best $\nu_{\min}$-variable to $\nu_{\max}$-variable regressor subsets, where

  $$1 \leq \nu_{\min} \leq \nu_{\max} \leq n.$$

- For example consider the index tracking of the UK FTSE index which comprises 100 stocks. The search for the best subset of stocks that can track the index might be restricted between the models comprising 20-70 stocks (variables). I.e. $\nu_{\min} = 20$ and $\nu_{\max} = 60$.

- ***Initially the variables could be sorted based on their significant values***.
- The order could be obtained by first fitting each variable separately in order to predict the response variable y (*as in the first step of the forward selection method*). Then, the variables are sorted in decreasing order based on their significant values. I.e. the most significant variables should be at the first position.
- This results the most significant variables to be included in the first generated submodels and thus, converge faster to the best submodels.

- The advantage of such search methods are:

  1. It does not require a complete search through all subset regression models. It avoids searching along unfavorable branches of the regression tree. This make it (computationally) feasible to investigate subsets comprising with more independent variables.

  2. It generates a reasonably small number of the best submodels and thus, simplifies their evaluation.

The House Prices data set – description

- SOURCE: *Long-Kogan Realty*, Chicago

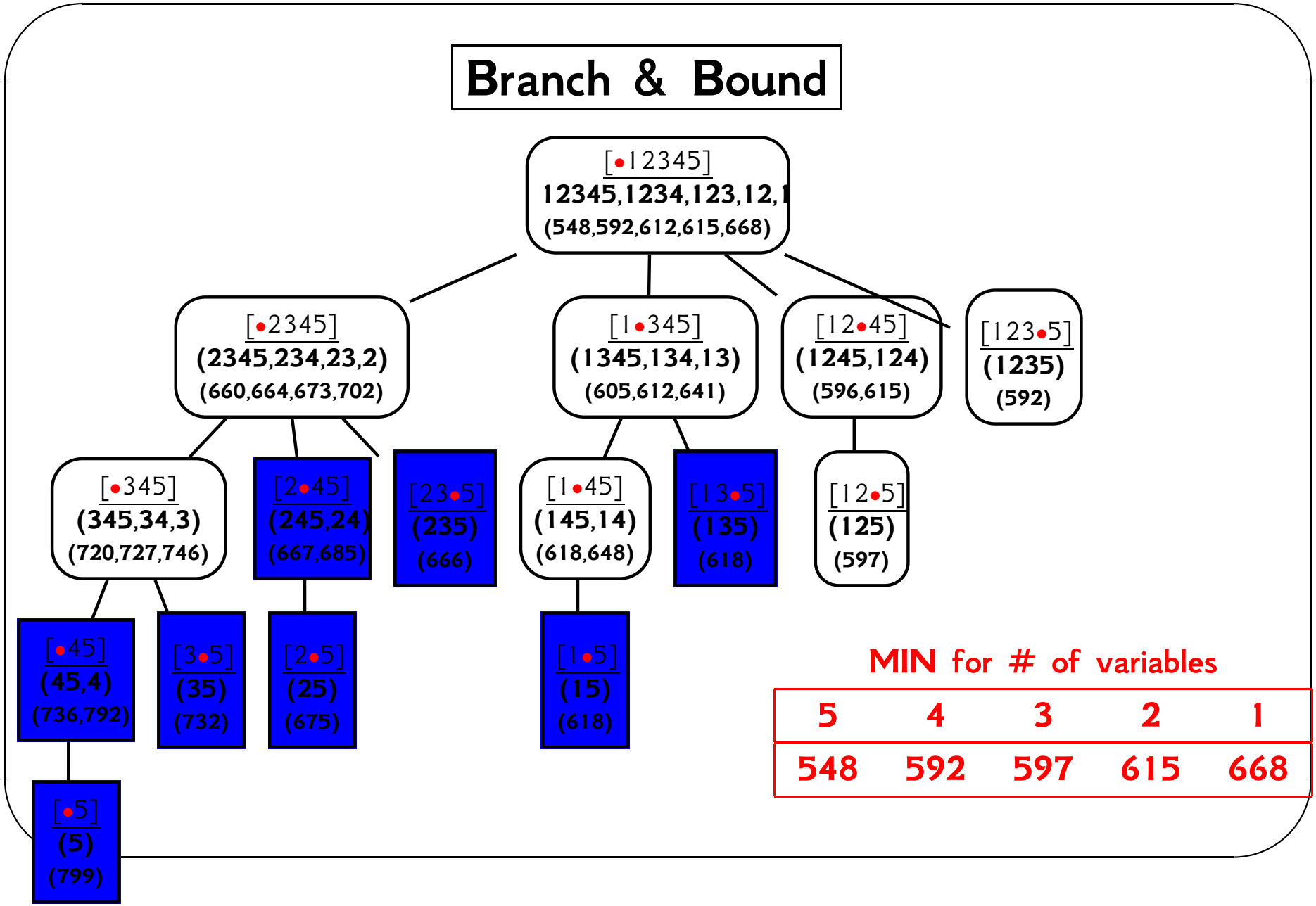
output		
	PRICE	Selling price of house in thousands of dollars
1	BDR	Number of bedrooms
2	FLR	Floor space in sq.ft.
3	FP	Number of fireplaces
4	RMS	Number of rooms
5	ST	Storm windows (1 if present, 0 if absent)
6	LOT	Front footage of lot in feet
7	TAX	Annual taxes
8	BTH	Number of bathrooms
9	CON	Construction (0 if frame, 1 if brick)
10	GAR	Garage size (0 = no garage, 1 = one-car garage, etc.)
11	CDN	Condition (1 = 'need work', 0 otherwise
12	L1	Location (L1 = 1 if property is in zone A, L1 = 0 otherwise)
13	L2	Location (L2 = 1 if property is in zone B, L2 = 0 otherwise)

The House Prices data set – selected models (without a constant term)

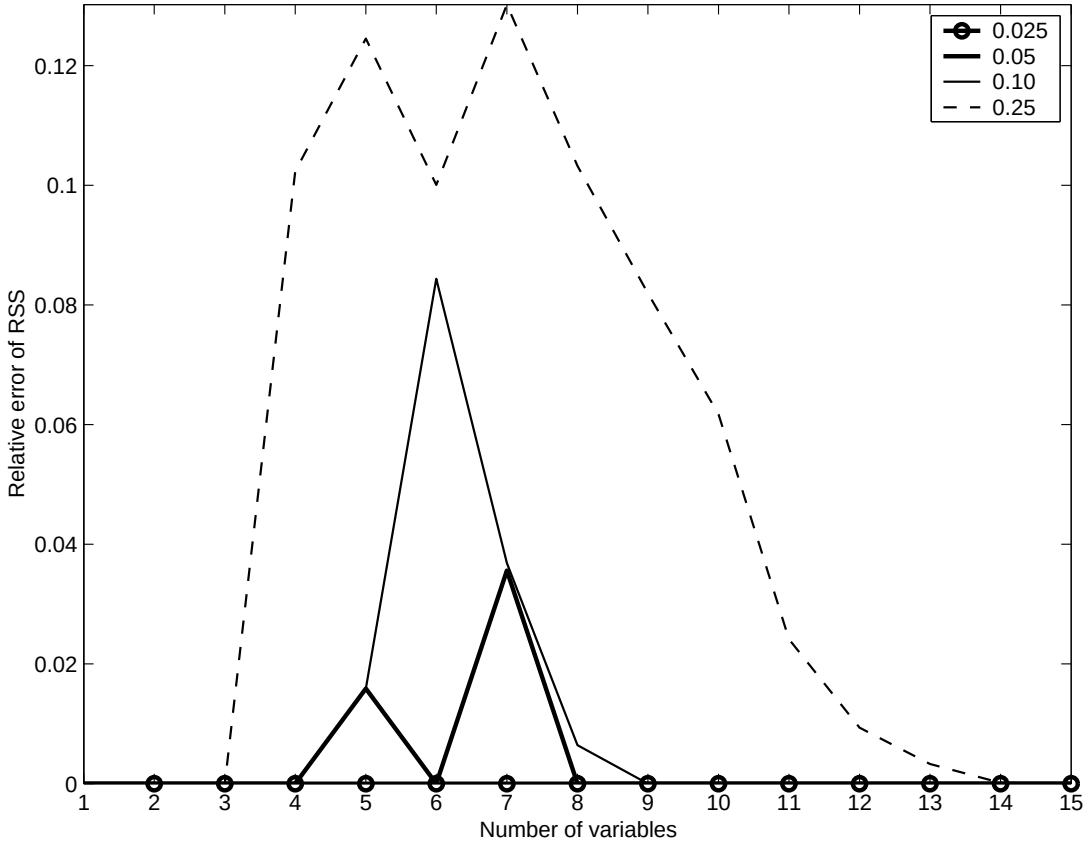
# of var.	R^2	Adjusted R^2	C_p	Model
Branch and Bound - exhaustive search				
0	0.00	-0.04	160.45	const.
1	0.54	0.50	62.56	FLR
2	0.67	0.63	40.04	FLR ST
3	0.76	0.71	26.38	FLR FP ST
4	0.81	0.76	18.94	BDR FLR FP ST
5	0.87	0.82	10.55	BDR FLR FP RMS ST
6	**0.90**	**0.86**	**6.20**	**FLR ST LOT CON GAR L2**
7	0.92	0.88	4.94	BDR FLR ST LOT CON GAR L2
8	0.93	0.89	4.81	BDR FLR RMS ST LOT CON GAR L2
9	0.94	0.89	5.96	BDR FLR FP RMS ST LOT CON GAR L2
10	0.94	0.89	7.51	BDR FLR FP RMS ST LOT BTH CON GAR L2
11	0.94	0.88	9.28	BDR FLR FP RMS ST LOT BTH CON GAR L1 L2
12	0.94	0.87	11.08	BDR FLR FP RMS ST LOT TAX BTH CON GAR L1 L2
13	0.94	0.86	13.00	BDR FLR FP RMS ST LOT TAX BTH CON GAR CDN L1 L2
Heuristic Branch and Bound (tolerance 10%) – differences				
10	0.94	0.89	7.77	BDR FLR FP RMS ST LOT CON GAR L1 L2
11	0.94	0.88	9.38	BDR FLR FP RMS ST LOT TAX CON GAR L1 L2

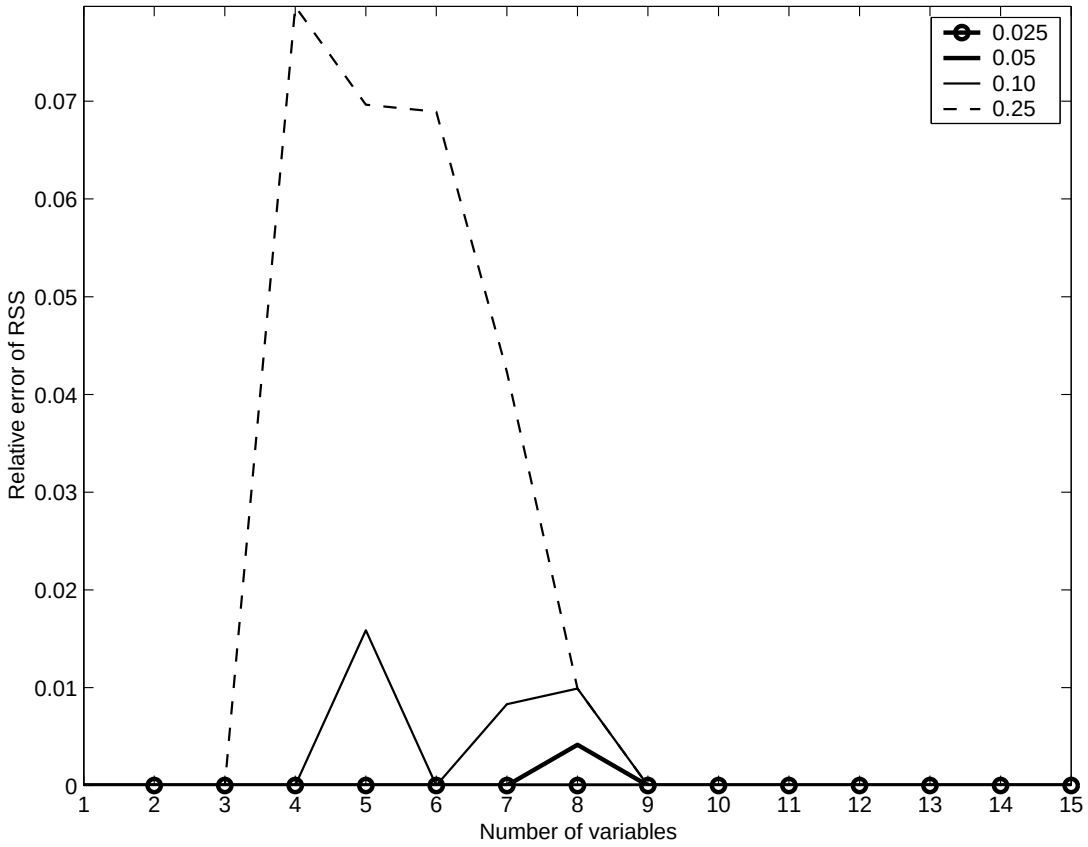
Search Strategies for the best submodels

- The most commonly used search method for generating the best subset models is called *Leaps and Bounds* (LB) [1].
- A recently developed *Branch & Bound* (BB) strategy outperforms the LB one [2,3].
- Heuristics based on the BB method in order to obtained models closed to the optimal ones and the investigation of sub-range models have been developed [4].


Branch & Bound
[•12345]
12345,1234,123,12,1
(548,592,612,615,668)
[•2345]
(2345,234,23,2)
(660,664,673,702)
[1•345]
(1345,134,13)
(605,612,641)
[12•45]
(1245,124)
(596,615)
[123•5]
(1235)
(592)
[•345]
(345,34,3)
(720,727,746)
[2•45]
(245,24)
(667,685)
[23•5]
(235)
(666)
[1•45]
(145,14)
(618,648)
[13•5]
(135)
(618)
[12•5]
(125)
(597)
[•45]
(45,4)
(736,792)
[3•5]
(35)
(732)
[2•5]
(25)
(675)
[1•5]
(15)
(618)
[•5]
(5)
(799)
MIN for # of variables
5 4 3 2 1
548 592 597 615 668

Experimental results using the Heuristic BB method




0.025
0.05
0.10
0.25
Relative error of RSS
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Number of variables

The Leaps & Bounds and Branch & Bounds methods

	Generating all models			Cutting sub-trees			
	Time (Sec.)			LBA		BBA	
# of Var	LBA	BBA	LBA / BBA	Time	Nodes	Time	Nodes
15	2.17	0.15	14	0.17	1958	0.02	969
16	4.55	0.28	16	0.15	1556	0.03	760
17	9.49	0.54	18	0.58	5932	0.08	2966
18	19.93	1.07	19	0.94	8870	0.13	4383
19	43.64	2.17	20	2.55	23316	0.29	11655
20	88.73	4.48	20	11.35	108806	1.07	54403
21	184.51	8.65	21	7.77	64348	0.80	32174
22	387.50	17.02	23	7.13	49994	0.72	24988
23	811.99	34.16	24	8.18	57060	0.81	28511
24	1739.70	68.20	26	62.76	454332	5.74	227159
25	3617.78	136.00	27	53.66	358008	5.60	178997
25	1h	2 min		1 min		6 sec	

Performance of the various versions of the BBA

The BB method with preordering of the variables is indicated as BB–1.

	Exhaustive methods		Heuristics with $\tau = 0.1$		Heuristics with $\tau = 0.25$	
# of Var	BB	BB–1	HBB	HBB–1	HBB	HBB–1
15	0.02	0.01	0.01	0.01	0.009	0.003
20	1.14	0.05	0.48	0.02	0.16	0.009
25	5.60	0.32	1.78	0.05	0.20	0.010
30	22.54	1.09	3.46	0.04	1.46	0.005
35	171.64	3.01	42.47	0.32	4.77	0.040
40	10049.32	45.09	168.17	1.31	12.27	0.030
41	3197.91	63.22	80.94	1.12	0.89	0.070
42	28176.72	76.09	4949.46	3.15	255.20	0.090
43	31567.22	289.52	1353.42	4.50	115.70	0.093
44	3806.57	89.07	266.99	2.78	11.93	0.086
45	47342.35	149.80	2105.87	1.74	17.25	0.042
45	13 h	2.5 min	36 min	2 sec	17 sec	0.042 sec